Design computationnel de protéines pour la prédiction de structure

Audrey Sedano-Pelzer

Laboratoire de Biochimie

École Doctorale de l'École Polytechnique

Design computationnel de protéines pour la prédiction de structure

THESE

présentée et soutenue publiquement le 23 Avril 2013

pour l'obtention du

Doctorat de l'École Polytechnique

(spécialité Bioinformatique)

par

Audrey Sedano

Composition du jury

Dr. Jean-François Gibrat	Rapporteur
Dr. Philippe Minard	Rapporteur
Dr. Isabelle André	Examinateur
Dr. Jean-Marc Steyaert	Co-directeur de thèse
Dr. Thomas Simonson	Directeur de thèse

Remerciements

Tout d'abord, je tiens à remercier les membres du jury, *Isabelle André*, et spécialement les rapporteurs, *Jean-François Gibrat* et *Philippe Minard*, qui n'ont pas rechigné à se plonger dans le manuscrit avec le sourire. Sans oublier les différents collaborateurs qui ont participé à cette thèse, *M. Valerio-Lepiniec*, *I. Guijarro*, *C. Malosse*, *E. Nicol*, *O. Robine*, *H. Sanejouand*. Je remercie également la *DGA* qui a financé cette thèse. Merci à *Michel et Nathalie S.* qui ont passé du temps à me préparer.

Je remercie *Thomas Simonson* de m'avoir ouvert la porte de la bioinformatique lorsque j'étais en Master d'Informatique, et de m'avoir formée à son domaine. J'ai beaucoup appris à travers sa rigueur scientifique, son obsession du détail et sa détermination à défendre ses idées. Je ne le remercierai jamais assez de m'avoir permis de réaliser ce que je voulais faire de ma vie. Je remercie également *Jean-Marc Steyaert*, qui même à travers les quelques feuilles de mon dossier de candidature a su me faire confiance et me donner une chance de rentrer au sein de cette École. Tout ce que j'ai entrepris et que j'entreprendrai dans l'avenir n'aura été possible que par son soutien et son accompagnement durant ces cinq ans. Je remercie enfin *Pierre Plateau* pour la qualité de son encadrement, sa gentillesse et sa disponibilité. Merci aussi à *Yves Méchulam* de m'avoir accueillie dans son laboratoire et de m'avoir permis d'être moniteur sur l'École. Merci aussi à *Thanh-Tâm Lê* et *Pierre Legrain* d'avoir été tant à l'écoute, dans les bons mais surtout les pires moments.

Bien sûr, je pense à *Emma* et à *Caroline* qui m'ont tout appris en biologie expérimentale, depuis le maniement de la pipette, en passant par le montage (parfois musclé) des colonnes de cobalt et l'observation des cristaux.

Mes remerciements vont ensuite évidemment à ceux qui m'ont entourée dans le laboratoire. Tout d'abord, *Thomas G.* et *David* qui auront été des aides et des soutiens précieux pour toute la partie bioinformatique. Pour leur patience, leur pédagogie, leur gentillesse et leur disponibilité. Ensuite, merci à *Pierre-Damien* (pi-di-ci) et à *Cédric* pour tous les conseils sur la thèse, la rédaction, les manips, la vie, le labo, et pour les pauses café, les discussions, les fous rires...

Je n'oublie pas non plus : *Titine* pour ta bonne humeur et tous tes conseils précieux, *Michou* pour ta gentillesse, *Marc D.* pour m'avoir fait peur dans les couloirs certains soirs tard, *Michel* pour tes conseils scientifiques, *Myriam* pour tes discussions scientifiques ou non, *les Pascaux* pour votre dévouement et votre soutien, *les deux Catherine* pour votre organisation, *Guillaume* pour toutes les petites choses que tu fais pour le labo ou pour nous faciliter la vie, *Alexey* pour tes bonjours discrets, *Laura* pour tes apparitions qui sentent bons le sud ensoleillé. Ainsi que *Etienne* pour le partage d'expérience bon comme mauvais, *Solène* pour nos deux ans de galères ensemble, *Auriane* pour les pauses, les rires, les blagues, et le coup de fil à ma place qu'on oubliera jamais (!). Sans oublier la nouvelle équipe, *Marc*, *Régis*, *Zeineb* et *Juliette* (à quand la dégustation de thé ?) qui ont donné une souffle nouveau et joyeux dans mon quotidien.

Une pensée particulière pour *Mélanie* pour toutes nos longues discussions dans ton bureau, ton amitié et ton soutien sans limite.

Je remercie aussi mes *professeurs* de l'Université de Toulon, qui ont cru en moi et m'ont soutenue jusqu'au bout.

Un énorme MERCI à mes *parents* sans qui je ne serai pas arrivée jusqu'ici. Pour leur amour et leur dévouement. Pour leur confiance et leur encouragement. Pour toutes ses valeurs qui me font être ce que je suis. Pour leur patience et leur soutien. À *Papa*, pour

m'avoir donnée le goût de la curiosité scientifique, le perfectionnisme et pour m'avoir permis de faire des études, de m'y avoir encouragée et soutenue. À *Maman*, pour m'avoir donnée le côté artiste, le sens des priorités et l'intelligence de la vie et qui m'a appris à toujours croire en mes rêves et à me battre pour y arriver. J'espère qu'ils sont fiers de moi aujourd'hui.

Un autre MERCI à mes *grands-parents*. Particulièrement, à *Papou* qui m'aura donné le goût de la biologie et de "bidouiller" sur l'ordinateur dès mon enfance, et qui, de là où il est, est sûrement fier de me voir ici. À *Minette*, qui m'a donnée tant d'amour toutes ses années. Et une pensée à *Grand-papy* : pour ses traces dans lesquelles j'ai marché, avec la science (et l'ombre de Marie Curie) et son violon.

Enfin, un énorme MERCI à mon mari, *Johan*, qui aura vécu de très très près, les pires et les meilleurs moments de ma thèse, au quotidien. Pour les pleurs et les fous rires, pour les raz-le-bol et les fiertés. Je m'en excuse et te redis merci. Sans ton soutien, ta compréhension, et ton amour je ne serai pas arrivée au bout. Et... "Ouf, c'est fini !"

A toi, H., qui m'aura soutenue dans les derniers jours de préparation de la soutenance, et qu'il l'aura vécu au plus près, avec moi...

Et enfin, aux *Shadok*, je les remercie de m'avoir inculquée une de mes plus grandes valeurs : *"En essayant continuellement, on finit par réussir. Donc plus ça rate, plus on a de chances que ça marche..."*, qui aura été appliquée scrupuleusement durant toute ma thèse...

Merci à tous...

À mes parents.

Table des matières

Liste des figures

Liste des tables

Abréviations

1CKA code PDB du domaine SH3 de la protéine Crk *mouse*

1CSK code PDB du domaine SH3 de la protéine Csk

1CKA-CORE protéine mutante du domaine SH3 de code PDB 1CKA, avec les positions du cœur hydrophobe non mutées

1CKA-LIG-V protéine mutante du domaine SH3 de code PDB 1CKA, avec les positions fonctionnelles non mutées et avec la mutation V170

1CKA-LIG-W protéine mutante du domaine SH3 de code PDB 1CKA, avec les positions fonctionnelles non mutées et sans mutation en W170

1CKA-*old* protéine mutante du domaine SH3 de code PDB 1CKA, avec des paramètres de génération de séquences antérieurs

1CKA-*WT* protéine sauvage du domaine SH3 de code PDB 1CKA

ADN Acide DésoxyriboNucléique

AMBER Assited Model Building and Energy Refinement, champ de force

Amp Ampicilline

BET Bromure d'éthidium

BLOSUM *BLOcks of amino acid SUbstitution Matrix*

BSA Albumine sérique bovine

CASA *Coulomb Accessible Surface Area*

CD Dichroïsme circulaire

CHARMM Chemistry at HARvard Macromolecular Mechanics, champ de force

CPD *Computational Protein Design* ou Design computationnel de protéine

Da Dalton, unité de masse atomique

DO Densité optique

DTT Dithiothréitol

E. coli *Escherichia coli*

EDTA Acide éthylène diamine tétraacétique

GASA *GB Accessible Surface Area*

GB *Generalized Born* ou Born généralisé

HEPES Acide 4-(2-hydroxyethyl)-1-piperazineethanesulfonique

HPLC Chromatographie Liquide Haute Performance

IPTG Isopropyl-b-D-thiogalactoside

kDa kiloDalton

MALDI *Matrix Assisted Laser Desoption Ionization*

MSA *Multiple Sequences Alignment* (Alignement multiple de séquences)

PAGE *PolyAcrylamid Gel Electrophoresis* (Gel d'électrophorèse)

PCR *Polymerase Chain Reaction*

PDB *Protein Data Bank*

PEG PolyEthylène Glycol

Pfam *Protein Family Databank*

PMSF *PhenylMethylSulfonyl Fluoride*

Proteus programme de simulation de conformation développé par Simonson *et al.*

PyMol *Python-based MOLecular visualization system*, programme de visualisation de structures chimiques en 3D

RMN Résonance Magnétique Nuléaire

RMSD *Root Mean Square Deviation* (écart quadratique moyen)

SD Séquence de Shine-Dalgarno

SDS *Sodium Dodécyl Sulfate*

SDS-PAGE Electrophorèse sur gel polyacrylamide en présence de SDS

SH3 ou SH2 *SRC Homology 3 domain* ou *SRC Homology 2 domain*

TBE Tris-Borate-EDTA

TEMED N,N,N,N'-tetra-methylethylenediamine

ToF *Time of Flight* ou temps de vol

Tris Tris(hydroxyméthyl)aminométhane

WebLogo application sur le web pour la représentation graphique (en logo) d'alignement multiple de séquences (fréquence d'apparition des acides aminés)

X-PLOR programme utilisé pour l'exploration de l'espace conformationnel des macromolécules

Introduction, état de l'art et contexte de recherche

Chapitre 1

Les protéines : Niveaux d'organisation et méthodes de prédiction

"[...] Il est des corps qui, métamorphosés une fois, conservent à jamais leur nouvelle forme ; mais il en est d'autres qui ont reçu du ciel le privilège de se transformer à leur gré. C'est le vôtre, divin Proteus."
Métamorphoses, Ovide

Les protéines sont une des plus importantes classes de molécules présentes dans tous les organismes vivants. En 1839, le chimiste hollandais G.J. Mulder publia des résultats sur l'analyse de la fibrine du sang, des albumines du sérum sanguin et de l'œuf. Ceux-ci indiquent que c'étaient des composés quaternaires (C,H,O,N) avec des pourcentages quasiment identiques pour ces quatre atomes et qui contenaient des traces variables de soufre et de phosphore. Sur la suggestion du chimiste suédois Berzelius, Mulder désigna ces composés sous le nom de protéines. Étymologiquement, le terme "protéine" vient du grec ancien *protos* qui signifie *premier*. Ceci fait référence au fait que les protéines sont les principaux composants des cellules. Beaucoup de fonctions cellulaires sont assurées par les protéines : organisation dans l'espace de la cellule (protéines de structure), transfert de molécules dans et en dehors des cellules (protéines de transport), modulation de l'activité

d'autres protéines (protéines régulatrices), transmission de signaux extérieurs (protéines de signalisation), ...

De l'ADN aux protéines Les protéines sont des chaînes d'acides aminés dont l'assemblage est gouverné par le code génétique. L'acide désoxyribonucléique (ADN) est une molécule, en forme de double hélice, située dans le noyau des cellules et qui contient l'information génétique, autrement dit l'ensemble des caractères s'exprimant dans un organisme. L'expression d'un gène est alors constituée de deux étapes. D'une part, la transcription qui est la copie d'une partie de l'information de l'ADN en ARN (acide ribonucléique). D'autre part, la traduction de l'information génétique qui fait intervenir les ribosomes, où l'enchaînement de codons de l'ARN est converti en acides aminés. Une protéine est donc un assemblage linéaire d'acides aminés ; on parle plus précisément de polypeptides pour des molécules ayant entre 2 et 50 acides aminés. Cette nature séquentielle des protéines a été mise en évidence pour la première fois par Frederick Sanger lors du séquençage de l'hormone insuline (prix Nobel de chimie en 1958).

Dans des conditions physiologiques, une séquence protéique ne reste pas sous forme d'un long filament non structuré mais comporte différents niveaux de structuration, allant de sa séquence en acides aminés à sa structure tridimensionnelle (3D). Sa structure 3D, la plupart du temps stable et bien déterminée, confère à la protéine une fonction biologique particulière. Une modification de la structure 3D, même minime, peut donc entraîner une diminution, voire la perte de l'activité de la protéine. Ainsi, la détermination de la structure tridimensionnelle de ces complexes peut aider à la compréhension des processus cellulaires.

Relation séquence-structure-fonction La fonction d'une protéine est intimement liée à sa forme tridimensionnelle. Autrement dit, comprendre le rôle d'une protéine nécessite la connaissance de sa structure. En 1954, Anfinsen (prix Nobel en 1972) énonça que toute l'information nécessaire au repliement d'une protéine dans sa structure fonc-

tionnelle (native) se trouve dans sa structure primaire, par conséquent dans sa séquence [Anfinsen *et al.* 1954 ; Anfinsen 1973]. Cette observation est depuis connue sous le nom de principe d'Anfinsen. D'après ce principe, il est thermodynamiquement possible d'accéder à la structure tertiaire d'une protéine à partir de sa seule séquence. Cette affirmation est satisfaisante pour envisager la modélisation de protéines, la description correcte de la séquence d'acides aminés étant en principe suffisante pour accéder à la structure tertiaire.

Dans le contexte actuel, alors que les programmes de séquençage des génomes fournissent d'importantes quantités de données relatives aux séquences d'ADN (donc de protéines) l'enjeu est de taille : pour de très nombreuses protéines, seule la séquence est connue. La détermination expérimentale de la structure (RMN, cristallographie aux rayons X ou microscopie électronique) est une tâche beaucoup plus longue, difficile et coûteuse que le séquençage. La prédiction théorique des structures 3D est donc devenue une nécessité pour compléter nos connaissances actuelles sur les génomes.

La structure 3D d'une protéine est une donnée complexe. Ce premier chapitre présente, dans une première partie, les différents niveaux de structuration des protéines. La deuxième partie présente les principales stratégies de prédiction de structure 3D de protéines à partir des séquences d'acides aminés. Ce qui nous amènera à nous poser le problème inverse de la prédiction de structure et à introduire les méthodes de design computationnel de protéines (ou *Protein Design* en anglais).

1.1 Structure des protéines

Les caractéristiques spatiales des protéines sont la clé de leurs fonctions. La structure des protéines peut être définie à plusieurs niveaux :

– Structure primaire d'une protéine : composition et séquence d'acides aminés, autrement dit sa formule chimique.

- Structure secondaire : structure locale qui rend compte de l'organisation spécifique d'un fragment d'acides aminés consécutifs dans l'espace.
- Structure tertiaire : structure tridimensionnelle, organisation spatiale complète des structures locales de la molécule
- Structure quaternaire : organisation de protéines oligomériques qui sont des assemblages non covalent de sous-unités, association de plusieurs protéines en un complexe (protéines multimériques).

1.1.1 Structure primaire

Les protéines sont des polymères dont la brique élémentaire est l'acide aminé (figure 1.1). Un acide aminé est une molécule composée d'un carbone asymétrique, le carbone C_α lié à une fonction amine NH_2, une fonction carboxyle $COOH$, un hydrogène H et une chaîne latérale ou radical R. Il existe 20 acides aminés principaux, qui diffèrent par la nature de leur chaîne latérale. La nature de ce radical confère aux acides aminés leurs propriétés physico-chimiques particulières. Les formules semi-développées des 20 acides aminés sont présentées dans la figure 1.2, et les codes à une et à trois lettres pour représenter les acides aminés sont listés dans le tableau 1.1.

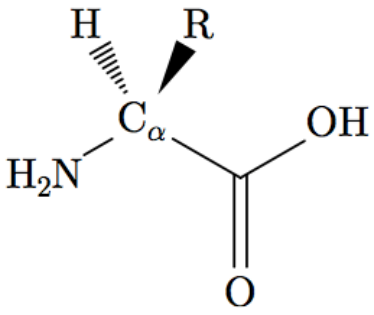


Figure 1.1 – Représentation générique des acides aminés (sauf pour la Proline, de forme cyclique).

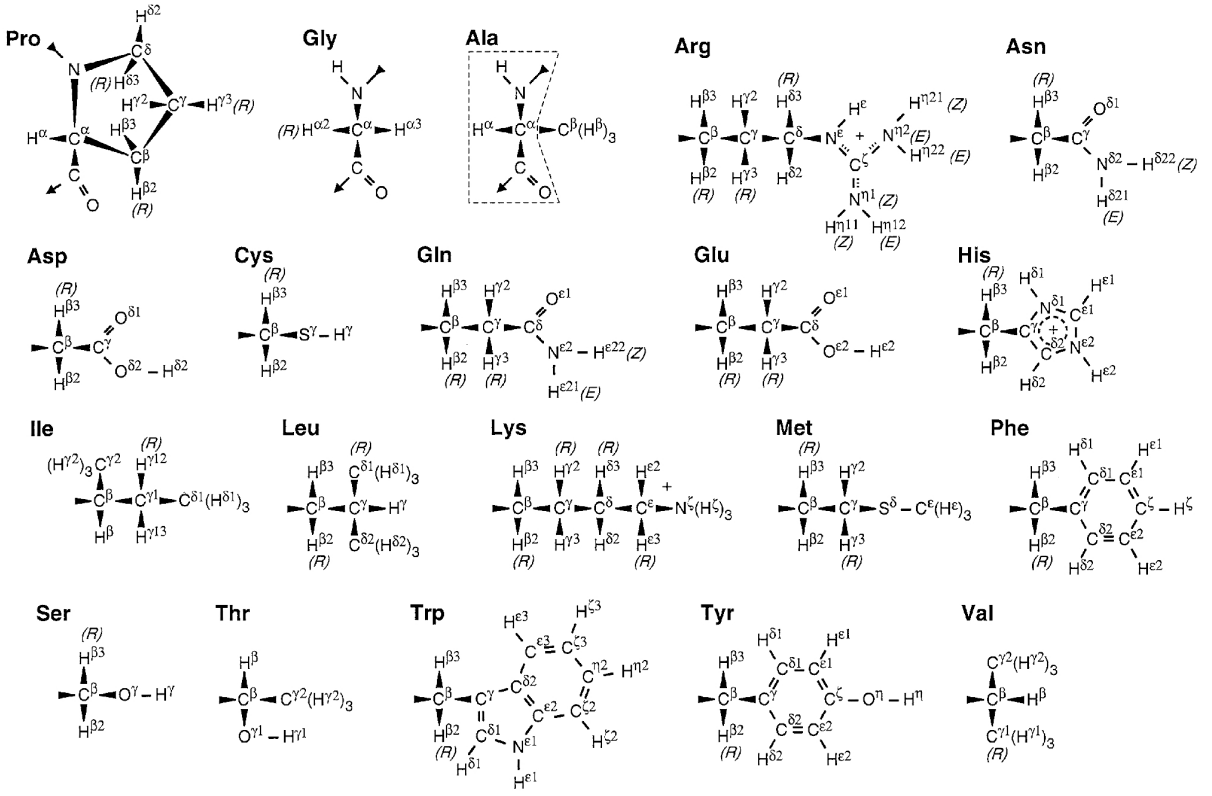


Figure 1.2 – Tableau récapitulatif des chaînes latérales des 20 acides aminés. On peut y voir le code à 3 lettres pour chaque acide aminé.

Nom	Abréviations	Nom	Abréviations
Alanine	Ala -A	Leucine	Leu - L
Arginine	Arg - R	Lysine	Lys - K
Asparagine	Asn - N	Méthionine	Met - M
Acide aspartique	Asp - D	Phénylalanine	Phe - F
Cystéine	Cys - C	Proline	Pro - P
Acide glutamique	Glu - E	Sérine	Ser - S
Glutamine	Gln - Q	Thréonine	Thr - T
Glycine	Gly - G	Tryptophane	Trp - W
Histidine	His - H	Tyrosine	Tyr - Y
Isoleucine	Ile - I	Valine	Val - V
Asp ou Asn	Asx - B	Glu ou Gln	Glx - Z
Inconnu	X		

Table 1.1 – Tableau récapitulatif des codes à une et trois lettres pour chaque acide aminé.

1.1.1.1 Classification des acides aminés

Il est intéressant de classer les acides aminés par groupe en prenant en compte leur nature et leurs propriétés physico-chimiques. Les acides aminés peuvent être qualifiés d'hydrophobes, d'après l'hydrophobicité (aversion pour l'eau) de leurs chaînes latérales, de polaires ou de chargés. Taylor [1986] en a ainsi proposé une classification assez fine, avec des recoupements entre classes, représentée par le diagramme dans la figure 1.3.

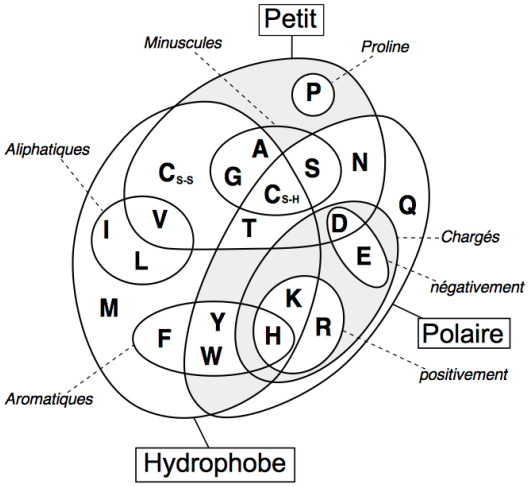


Figure 1.3 – Diagramme de Venn représentant la classification des acides aminés d'après Taylor.

Cette classification donne un aperçu de la complexité du problème qui consiste à créer des groupes. Par exemple, le Tryptophane peut être considéré comme hydrophobe (en raison de son cycle aromatique) ou polaire (en raison de son atome d'azote). L'Histidine peut être ou non chargée positivement selon le pH, d'où son inclusion dans deux groupes. La Cystéine est un cas particulier : au sein des structures 3D, deux Cystéines proches peuvent s'apparier pour former une liaison covalente appelée pont disulfure. Les Cystéines ainsi appariées ne sont plus polaires. Ces ponts disulfures sont, par ailleurs, particulièrement importants pour le maintien des structures 3D.

Même si nous avons utilisé ce genre de classification d'acides aminés dans ce travail de thèse, il existe d'autres méthodes pour les regrouper. Par exemple, la propriété d'hydrophobicité est probablement mieux représentée par l'utilisation d'échelles d'hydrophobicité

que par la constitution de classes. Plusieurs dizaines d'indices d'hydrophobicité ont été publiés à ce jour [Cornette *et al.* 1987 ; Biou *et al.* 1988 ; Kawashima *et al.* 1999]. Autre exemple, la taille de la chaîne latérale peut permettre d'établir des groupes [Taylor 1986].

Toutes ces classifications peuvent être très utiles pour analyser des alignements multiples de séquences, en minimisant la variabilité et les petites différences entre acides aminés. Néanmoins, pour quantifier les différences, il existe des matrices de score de substitution ou de similarité.

1.1.1.2 Substitutions d'acides aminés

Les matrices de substitutions sont sans doute les outils les plus utilisés dans les analyses d'alignements multiples. En effet, elles permettent de mesurer à quel point un acide aminé peut être similaire à un autre. La matrice de substitution est écrite en fonction des acides aminés et est de taille 20×20.

Un des modèles les plus simples considère que les taux d'échanges sont les mêmes quels que soient les deux acides aminés : la matrice de substitution du modèle F81 [Felsenstein 1981]. Cependant, il existe *in vivo* une sélection purificatrice ou de correction extrêmement forte au niveau de la protéine, pour qu'elle puisse conserver sa fonction biologique à tout prix. Si une telle contrainte parait évidente, la façon de la modéliser l'est beaucoup moins : quel modèle permettrait le mieux de capter les effets de cette sélection au niveau protéique ?

La première approche [Dayhoff *et al.* 1972], consiste à dire que les mutations conservant les propriétés biochimiques des acides aminés sont les moins délétères et donc les plus susceptibles d'être fixées. Par exemple, le remplacement d'une Valine par une Alanine serait moins problématique pour la protéine que le remplacement de cette même Valine par une Arginine. Il s'agit donc de construire une matrice de substitution 20×20 telle que les substitutions conservatrices soient les plus fréquentes. En pratique, ces paramètres sont heuristiques. La première matrice [Dayhoff *et al.* 1972] utilisait des alignements de

séquences afin d'estimer les paramètres de la matrice 20×20, en comptant le nombre de fois où des substitutions avaient été observées entre des séquences ayant 85% de similarité. Ils supposaient qu'avec 85% de similarité, les différences entre les deux séquences étaient dues à des substitutions simples et non à des substitutions multiples, c'est à dire des substitutions de a vers b via d'autres acides aminés [Jones *et al.* 1992]. Toutefois, si de telles matrices permettent de prendre en compte une sélection au niveau de la protéine, elles présentent le désavantage de représenter de la même manière les substitutions d'un acide aminé à un autre, quel que soit le site considéré. Or les contraintes liées à la sélection naturelle ne sont pas les mêmes sur tous les sites de la protéine, et elles sont au contraire fortement hétérogènes : un acide aminé appartenant au site fonctionnel de la protéine n'a pas les mêmes propensions substitutionnelles qu'un acide aminé dont le seul rôle serait d'être à la surface de la protéine.

1.1.1.3 Liaison peptidique et angles dièdres

Une séquence de protéine est formée par un enchaînement d'acides aminés, selon un ordre défini par la séquence d'ADN du gène correspondant. Une protéine de taille moyenne est constituée d'environ 100 à 300 acides aminés. La polymérisation de la chaîne protéique se fait par la perte d'une molécule d'eau lors de la condensation d'un groupement carboxyle α-COOH avec le groupement amine α-NH_2 du résidu suivant, comme on peut le voir dans le schéma 1.4. La liaison ainsi formée est dénommée liaison peptidique et l'acide aminé ainsi incorporé à la chaîne est dénommé résidu. Les atomes participant à la liaison peptidique forment le squelette de la protéine ou encore la chaîne principale, excepté le Cα.

La liaison peptidique est rigide en raison de la délocalisation des électrons du groupe carboxamide. Les liaisons $C = O$ et $N - H$ sont donc maintenues dans le même plan. La grande majorité des liaisons peptidiques dans les protéines sont de type *trans* : les groupes CO et NH pointent dans des directions opposées. De même, les longueurs des

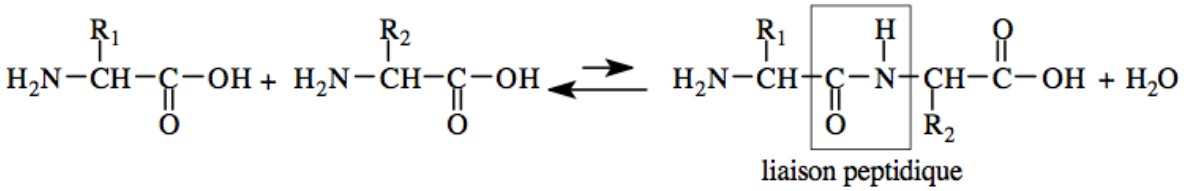


Figure 1.4 – Polymérisation de la chaîne protéique lors de la condensation du groupement carboxyle α-COOH avec le groupement amine α-NH_2 du résidu suivant.

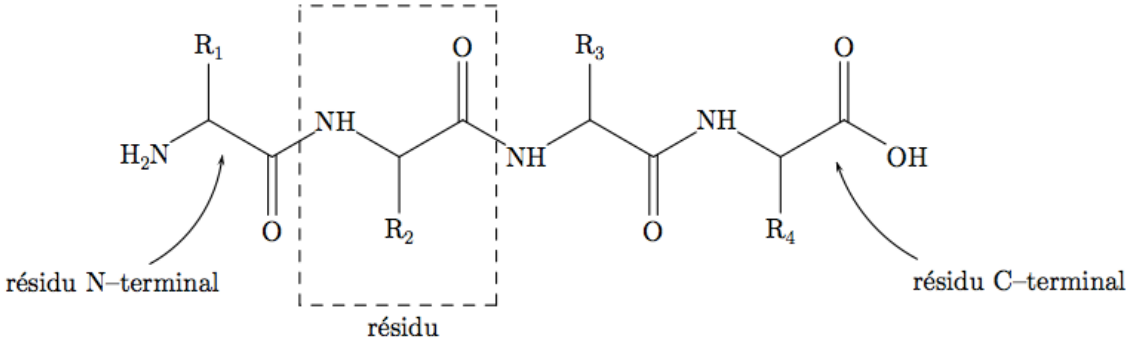


Figure 1.5 – Résidus d'un tétrapeptide avec les extrémités N-terminale et C-terminale.

liaisons chimiques varient très peu autour de leurs valeurs de référence. Les degrés de liberté pour le repliement de la chaîne protéique correspondent aux rotations autour des liaisons $NH - C_\alpha$ et $C_\alpha - CO$. Ces deux rotations sont décrites respectivement par les angles dièdres Φ et Ψ (schéma 1.6).

Les valeurs possibles de Φ et Ψ sont contraintes par l'encombrement stérique : certaines conformations sont impossibles car les atomes ne peuvent pas s'interpénétrer. Les valeurs permises pour Φ et Ψ ont été étudiées par Ramachandran, en modélisant les atomes par des sphères pleines, et en n'utilisant que des contraintes géométriques. Les résultats sont visualisés dans un plan Φ-Ψ, appelé diagramme de Ramachandran.

1.1.1.4 Diagramme de Ramachandran

Le diagramme de Ramachandran est une représentation graphique permettant d'analyser la conformation du squelette polypeptidique des protéines. Pour chaque acide aminé

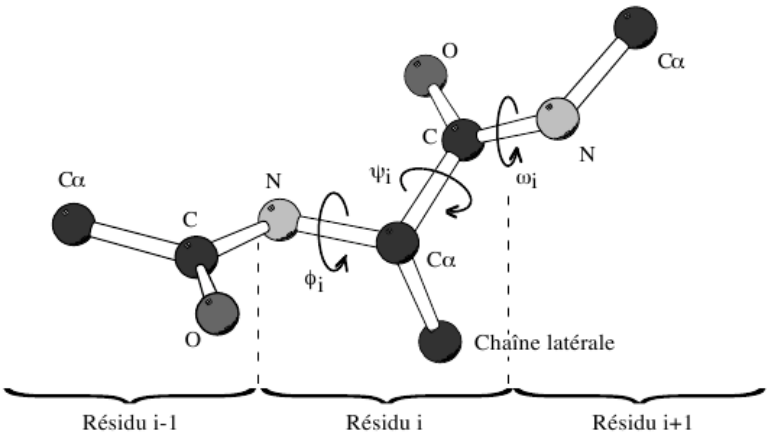


Figure 1.6 – Les angles dièdres permettant de décrire la structure secondaire d'une protéine. On y voit qu'un seul atome de la chaîne latérale pour le résidu i.

de la protéine, on porte la valeur de l'angle diédral Φ en abscisse et celle de l'angle dièdral Ψ en ordonnée, pour des valeurs de -180 à +180 degrés. Toutes les valeurs des angles Φ et Ψ ne sont pas possibles car certaines conduisent à des contacts trop proches entre atomes qui sont énergétiquement très défavorables. En effet, les volumes des chaînes latérales limitent considérablement les valeurs effectivement accessibles aux angles Φ et Ψ. Une étude systématique des combinaisons admissibles d'angles Φ et Ψ a été réalisée par le biologiste et physicien indien Ramachandran & Sasisekharan [1968].

Le diagramme de Ramachandran est depuis couramment utilisé pour analyser la structure des protéines. Certaines zones du diagramme sont très fortement défavorables. La Proline et la Glycine ont des localisations particulières. La Glycine a accès à un plus grand nombre de conformations que les autres résidus, en raison de sa chaîne latérale très réduite. Au contraire, dans le cas de la Proline, l'inclusion de l'azote de la chaîne principale dans le cycle de la chaîne latérale introduit une contrainte supplémentaire et limite davantage les valeurs prises par les angles Ψ.

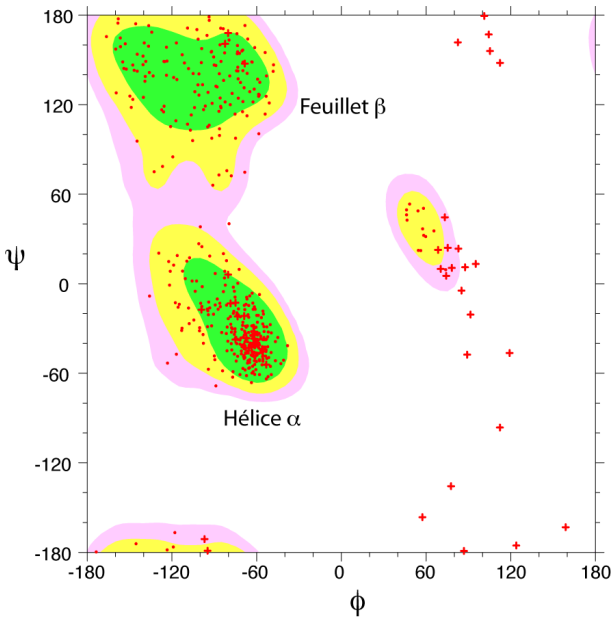


Figure 1.7 – Diagramme de Ramachandran pour une protéine. Les zones énergétiquement favorables sont représentées par des contours colorés. Chaque acide aminé est représenté par un point rouge. Les croix rouges correspondent aux acides aminés Glycine, qui ne comportent pas de chaînes latérales.

1.1.2 Structure secondaire

La structure secondaire désigne la conformation adoptée par la chaîne principale au niveau local. L'existence de l'hélice α et du feuillet β, qui sont les deux principales structures secondaires régulières, a été prédite par Pauling & Corey [1951]. Pauling et Corey recherchaient alors des structures locales régulières permettant de former le maximum de liaisons hydrogène, liaisons chimiques de faible énergie, au sein du squelette, tout en respectant les longueurs et angles de liaison connus [Pauling *et al.* 1951].

Nous avons vu précédemment que les liaisons N-Cα et Cα-C effectuaient librement des mouvements de rotation, mais seules quelques conformations parviennent à minimiser la gêne stérique. On peut répartir ces structures en catégories représentées dans le tableau 1.2.

Conformation	Φ (°)	Ψ (°)	Résidus/tour	Translation/résidu (Å)
Hélice α droite	-57	-47	3,6	1,50
Hélice α gauche	+57	+47	3,6	1,50
Hélice 3_{10} droite	-49	-26	3,0	2,50
Hélice π droite	-57	-70	4,4	1,15
Hélice gauche du collagène	-51	+153	3,0	3,13
Feuillet plissé β antiparallèle	-139	+135	2,0	3,40
Feuillet plissé β parallèle	-119	+113	2,0	3,20
Chaîne étirée	+/-180	+/-180		

Table 1.2 – Paramètres angulaires des différentes structures secondaires. La 4^e colonne représente le nombre de résidus nécessaires pour faire un tour, autrement dit l'inclinaison d'un résidu. La dernière colonne permet de connaître la longueur du peptide suivant la structure secondaire considérée et le nombre de résidus.

1.1.2.1 L'hélice alpha

C'est une structure locale répétitive et relativement compacte dont les caractéristiques principales sont :

- la structure est stabilisée par les liaisons hydrogène de l'atome d'oxygène C=O d'une liaison peptidique i avec l'atome d'hydrogène N-H de la liaison peptidique $i + 4$
- les radicaux des résidus sont à l'extérieur de l'hélice, ce qui minimise les encombrements stériques
- l'hélice ne peut s'établir qu'avec des résidus de la même série (L ou D)
- la configuration L des acides aminés privilégie un enroulement à droite (dans un enroulement à gauche, les chaînes latérales recouvrent trop le squelette de l'hélice)
- les plans des liaisons peptidiques sont parallèles à l'axe de l'hélice et forment le squelette de l'hélice.

On peut préciser en détail les caractéristiques de l'hélice de la façon suivante :

- pas : 0,54 nm par tour
- nombre de résidus par tour : 3,6
- translation par résidu : 0,15 nm
- diamètre de l'hélice : 0,50 nm
- angles dièdres : $\phi = -57\ ^\circ$ et $\psi = -47\ ^\circ$

– la liaison hydrogène C=O–H–N (d'une longueur de 0,286 nm) est presque parallèle à l'axe de l'hélice

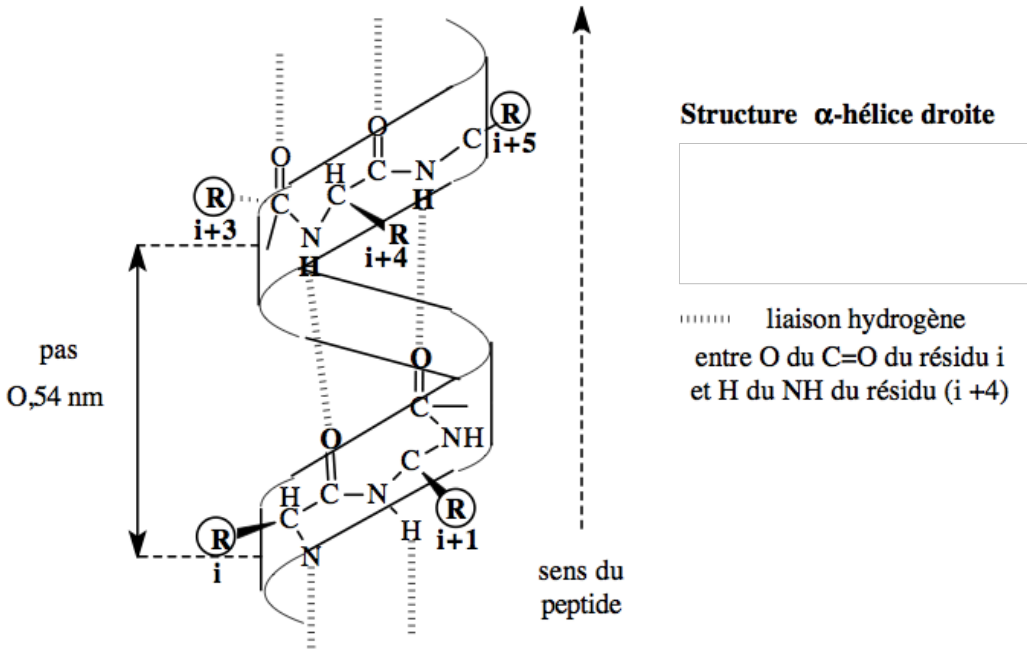


Figure 1.8 – Représentation schématisée d'une hélice α

Cette structure est favorisée par les résidus dont les chaînes latérales ne portent pas de charges et dont l'encombrement stérique est faible. Certains résidus déstabilisent l'hélice par la présence de charge dans leurs chaînes latérales (Acide aspartique, Acide glutamique, Arginine et Lysine). La Proline est un point de rupture d'une hélice pour deux raisons : il n'existe pas de N-H pour une liaison hydrogène, et le cycle rigide pyrrolidone engagé dans le liaison peptidique bloque la rotation, détruisant la continuité de l'hélice. L'hélice α est la structure secondaire la plus abondante puisqu'elle concerne environ 30 % des résidus en moyenne.

1.1.2.2 Le feuillet beta

Le feuillet β est stabilisé par des liaisons hydrogène entre des résidus éloignés le long de la séquence, dans des portions de chaîne en conformation étendue, les brins β. Deux

catégories de feuillets sont distinguées, selon l'orientation relative des brins : feuillets parallèles et feuillets antiparallèles. Les feuillets β parallèles et anti-parallèles ont des caractéristiques très similaires :

- parallèle : translation 0,32 nm et angles dièdres $\phi = -119\,^{\circ}$ et $\psi = 135\,^{\circ}$
- anti-parallèle : translation 0,34 nm et angles dièdres $\phi = -139\,^{\circ}$ et $\psi = 135\,^{\circ}$

Les chaînes latérales pointent alternativement d'un côté et de l'autre du feuillet. La nature non-locale des liaisons hydrogène dans les feuillets β les différencie fondamentalement des hélices α. Des brins β isolés, existant de manière stable hors des feuillets, ont également été décrits [Eswar *et al.* 2003]. Environ 20 % des résidus des protéines sont impliqués dans des feuillets β.

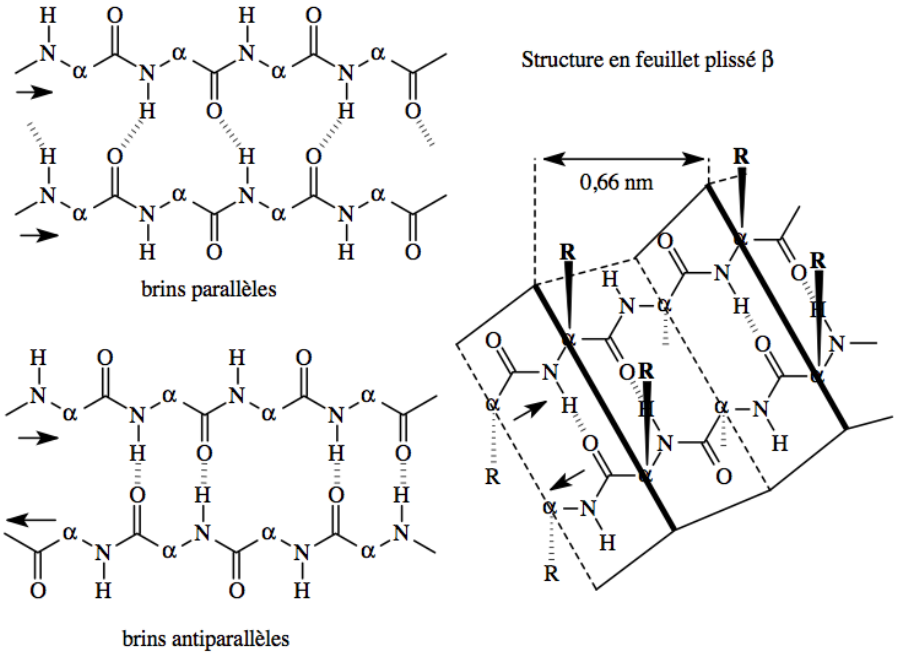


Figure 1.9 – Représentation schématisée d'un feuillet β

1.1.2.3 Les coudes

Certaines régions protéiques ne sont pas structurées dans des conformations périodiques. Toutefois leurs structures sont semblables par le fait qu'elles imposent un change-

ment brusque de direction de 180 ° : on les appelle coude ou tour β (β-*turn* ou *coil*). La lettre β rappelle qu'ils sont indispensables pour des feuillets de chaînes anti-parallèles. Une description plus fine de la boucle fait apparaître d'autres motifs réguliers, par exemple :

- Le coude β est formé de 4 résidus, stabilisés ou non par une liaison hydrogène [Lewis *et al.* 1971 ; Hutchinson & Thornton 1994]. Cette structure courte permet un changement de direction de la chaîne principale et concerne 25 % des résidus [Kabsch & SAnder 1983].
- L'hélice 3-10 est une hélice droite stabilisée par des liaisons hydrogène de type (i, $i+3$) [Donohue 1953]. Majoritairement courtes, les hélices 3-10 sont fréquemment rencontrées aux extrémités des hélices α et concernent 3 à 4 % des résidus [Barlow & Thornton 1988].
- L'hélice π est une hélice droite formée par des liaisons hydrogène de type (i, $i+5$) [Low & Baybutt 1952]. Elle concerne moins de 1 % des résidus [Fodje & Al-Karadaghi 2002].

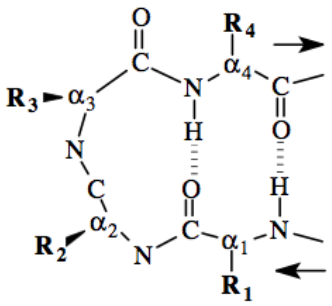


Figure 1.10 – Représentation schématisée d'un coude de 4 résidus.

Même avec cette classification enrichie, un grand nombre de résidus restent non décrits.

Les structures secondaires sont aujourd'hui largement utilisées dans un grand nombre d'applications en biologie structurale, car elles permettent une description simple et intuitive des structures 3D. La description des structures 3D en terme de structures secondaires est ainsi mise à profit, par exemple, en comparaison et en classification des structures.

Les logiciels de visualisation de structures proposent systématiquement une représentation des structures 3D mettant en évidence les structures secondaires : hélice α, feuillet β et éventuellement coudes.

1.1.2.4 Pelote statistique

Certaines régions protéiques ne sont pas structurées dans des conformations périodiques comme l'hélice α ou le feuillet β plissé : leur forme irrégulière est qualifiée de pelote statistique (*random coil*). Cette structure n'est pas pour autant inorganisée, elle obéit aux contraintes locales de voisinage.

1.1.2.5 Irrégularités des structures secondaires

L'observation des structures secondaires périodiques dans les structures disponibles fait apparaître un grand nombre d'irrégularités. Par exemple, la grande majorité des hélices α ne sont pas parfaitement linéaires, mais courbées à des degrés variés, voir coudées [Barlow & Thornton 1988 ; Kumar & Bansal 1998]. Des π-*bulges*, provoqués par des liaisons hydrogène de type $(i, i+5)$ dans les hélices α, ont également été décrits [Cartailler & Luecke 2004]. Les feuillets β comportent des altérations dans la régularité de l'appariement des brins, introduites par l'insertion d'un résidu dans l'un des brins : les β-*bulges* [Richardson *et al.* 1978]. Les β-*bulges* sont relativement fréquents dans les feuillets antiparallèles [Richardson & Richardson 2002] : l'étude de Chan *et al.* [1993] dénombre ainsi deux β-*bulges* par protéine. Ces β-*bulges* introduisent une discontinuité dans l'alternance d'orientation des chaînes latérales d'un côté et de l'autre du feuillet. Il a été suggéré que les β-*bulges* permettent d'accommoder une insertion dans la séquence sans perturber l'architecture globale du feuillet et pourraient également permettre d'éviter l'appariement indésirable avec un brin situé en bordure de feuillet [Richardson & Richardson 2002].

Les motifs de structures secondaires sont donc souvent sensiblement différents des modèles théoriques proposés par Pauling et Corey. Ces irrégularités peuvent, dans certains cas, rendre délicate la détection automatique des structures secondaires.

1.1.3 Structure tertiaire

L'arrangement spatial des structures secondaires locales aboutit à une forme globale spécifique de la protéine maintenue par des interactions qui peuvent être de natures différentes :

- liaison covalente : pont disulfure
- liaisons ioniques entre groupements chargés de signes opposés (pont salin)
- interactions électrostatiques entre dipôles permanents et groupements ionisés ou encore entre deux dipôles : les plus répandues sont les liaisons hydrogène
- attractions hydrophobes (entre groupes apolaires subissant des forces de répulsion par l'eau, ce qui favorise leur rapprochement)
- interactions des chaînes latérales des résidus avec le solvant : dans l'eau, les chaînes latérales polaires pourront être exposées au solvant, alors que les chaînes latérales apolaires auront tendance à "s'enfouir" dans des poches hydrophobes de la protéine.

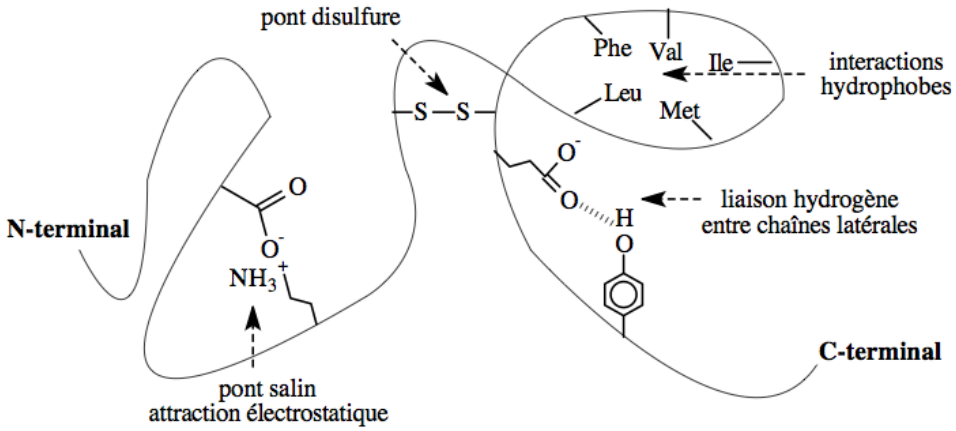


Figure 1.11 – Liaisons ou interactions entre chaînes latérales des résidus impliquées dans la structure tertiaire des protéines.

Méthode expérimentale	Protéines	Complexes protéine/ADN	Total
Cristallographie aux rayons X	68 368	3 464	71 832
RMN	8 337	188	8 525
Microscopie électronique	305	120	425
Autres	185	7	192
Total	77 195	3 779	80 974

Table 1.3 – Nombre de structures connues dans la banque de structures PDB.

La structure tertiaire d'une protéine est le paramètre fondamental dont dépend l'expression de ses fonctions biologiques qu'elles soient structurales ou dynamiques. Cette structure est très dépendante des interactions que nous venons de décrire. Ces interactions subissent l'influence du milieu dans lequel elles se trouvent (solvant, température, pH, force ionique, agents destructeurs, ...). La conformation native de la protéine est la structure tertiaire qui correspond à celle qui exprime sa fonction biologique dans les conditions physiologiques. Un traitement détruisant cette structure génère la perte de sa fonction ; c'est la dénaturation. Elle entraîne un état plus désorganisé de la molécule.

L'ensemble des structures disponibles déterminées expérimentalement est regroupé dans la *Protein Data Bank* (PDB), http ://www.pdb.org/ [Berman *et al.* 2000]. Un fichier PDB contient l'information de structure sous la forme des coordonnées de tous les atomes de la protéine dans l'espace à trois dimensions. Les atomes d'hydrogène ne sont en général pas décrits dans les structures cristallographiques. Certains atomes de la protéine peuvent ne pas être décrits si la résolution ne le permet pas.

La PDB comporte à ce jour plus de 80 000 structures protéiques déterminées par cristallographie aux rayons X, par résonance magnétique nucléaire (RMN) ou par microscopie électronique (comme on peut le voir dans la tableau 1.3). De même que les banques de séquences, la PDB est redondante : elle contient de nombreuses structures de protéines dont les séquences sont très similaires, voire identiques (cas de protéines co-cristallisées avec différents inhibiteurs, par exemple), et des protéines mutantes.

1.1.3.1 Propriété d'homologie des protéines

Outre leur importance fonctionnelle, les structures secondaires et tertiaires présentent un intérêt cognitif majeur car elles sont mieux conservées que les séquences au cours de l'évolution. Deux protéines sont dites homologues si elles dérivent d'un ancêtre moléculaire commun. Au cours de l'évolution, des mutations s'opèrent sur les séquences d'ADN. Ces mutations sont conservées si les protéines codées sur les gènes conservent leur fonction et donc leur structure tridimensionnelle, en raison de la pression de sélection qui tend à maintenir la fonction. En conséquence, des séquences différentes peuvent adopter la même structure. L'homologie est généralement mise en évidence par la comparaison de séquences, quand l'ancêtre commun est encore assez proche. Si les séquences ont trop divergé, la comparaison de séquences ne permet pas de détecter l'homologie.

1.1.4 Structure quaternaire

La structure quaternaire des protéines regroupe l'association d'au moins deux chaînes polypeptidiques, identiques ou différentes, par des liaisons non-covalentes (liaisons hydrogène, liaisons ioniques, interactions hydrophobes), plus rarement des ponts disulfures. Chacune de ces chaînes est appelée monomère (ou sous-unité) et l'ensemble oligomère ou protéine multimérique.

L'effet hydrophobe est le facteur prépondérant dans l'assemblage des éléments structuraux, y compris dans l'association des sous-unités. L'hémoglobine est un exemple de structure quaternaire constituée de 4 sous-unités : 2 sous-unités α et 2 sous-unités β.

Ces assemblages polypeptidiques peuvent être parfaitement bien définis, mais certains de ces complexes peuvent seulement être transitoires tout en étant fonctionnellement importants. En effet, dans l'environnement cellulaire, les protéines sont rarement isolées, et les interactions qu'elles peuvent créer avec d'autres partenaires sont essentielles à leur fonction. L'étude de la structure quaternaire d'une protéine est donc aussi très importante pour bien comprendre la fonction biologique.

1.2 Familles de séquences

L'accroissement exponentiel des données protéiques issues des projets de séquençage de génomes complets aboutit à un goulot d'étranglement [Benson *et al.* 2009]. Identifier la fonction d'autant de protéines par des expérimentations biologiques (étude de leur structure 3D, de leurs interactions, etc.) n'est simplement pas réalisable. On ne dispose que de séquences primaires dont il faut tirer le maximum d'informations. Le recours actuel pour traiter ces données consiste à utiliser des méthodes bioinformatiques pour prédire une annotation fonctionnelle rapide des protéines récemment séquencées. L'annotation fonctionnelle des protéines passe alors par l'identification de groupes de séquences, plus communément appelés familles. Les familles visent à regrouper les protéines homologues, c'est-à-dire qui partagent une histoire évolutive commune. Les séquences d'une même famille sont souvent proches et possèdent donc des propriétés semblables. L'intérêt d'une telle classification est de permettre le transfert d'annotations. Lors de l'étude d'une nouvelle protéine (encore non-annotée), elle est assignée à une famille afin de transférer à cette protéine les connaissances acquises sur les autres protéines de la famille (issues de précédentes études). Le regroupement en familles des protéines peut se faire selon la similarité des séquences ou des structures, ou encore la proximité des compositions en domaines protéiques. Des méthodes récentes proposent de s'appuyer sur des combinaisons de ces critères avec des données biologiques telles que les voies métaboliques ou les profils d'expression [Hahne *et al.* 2008 ; Brehelin *et al.* 2010].

1.3 Prédiction de la structure des protéines : état de l'art

Les méthodes expérimentales telles que la cristallographie aux rayons X, la spectroscopie RMN et la microscopie électronique sont des techniques coûteuses et très difficiles à

mettre en place. De plus, le fossé entre le nombre de protéines séquencées et les structures 3D connues se creuse. Par conséquent, les modèles structuraux des protéines sont très importants pour prédire la structure tertiaire des protéines. Certaines études sur l'alignement structural des protéines de la banque PDB soulignent une forte probabilité pour qu'une séquence protéique ait déjà des types de repliement connus.

Comment déterminer la structure d'une protéine à partir de sa seule séquence ? Pour répondre à cette question, deux approches différentes sont couramment utilisées. La première approche consiste à construire un modèle 3D d'une protéine directement à partir des acides aminés de sa séquence. Les méthodes utilisant cette approche sont dites *ab initio* ou *de novo*. La deuxième approche consiste à essayer de trouver des protéines "similaires" dans des banques de protéines connues. A partir de ces protéines, un modèle 3D de la structure est construit.

1.3.1 Méthodes de prédiction *ab initio*

1.3.1.1 Méthodes *ab initio* et *de novo*

Jusqu'à présent, les méthodes décrites tentaient d'adapter la séquence à prédire sur des structures de la banque PDB. Les méthodes de prédiction *ab initio* ont pour but de proposer des structures 3D dans les cas où la modélisation par homologie et la reconnaissance de repliement ne fournissent pas de solution, ou parce que la séquence à prédire adopte une structure qui n'a jamais été observée. La modélisation *ab initio* consiste à rechercher la structure de la séquence cible en partant uniquement de la séquence, alors que les approches *de novo* utilisent les structures de la banque PDB pour extraire des fragments, qui sont ensuite assemblés pour construire un modèle.

La prédiction *ab initio* consiste à construire un modèle physique simplifié de la séquence et à effectuer une recherche exhaustive de l'espace conformationnel pour obtenir la structure de moindre énergie. Malgré les progrès accomplis [Hardin *et al.* 2002], la prédiction *ab initio* reste très difficile. L'espace conformationnel à explorer est immense et la

fonction d'énergie demeure particulièrement complexe à formuler. Néanmoins, ces modèles s'affinent de plus en plus. On peut citer l'optimisation de la définition de l'espace conformationnel par l'utilisation de librairies de rotamères (comme les utilise le programme SCWRL [Dunbrack & Karplus 1993a]).

La stratégie *de novo*, plus récente, tente de prédire la structure de petits fragments de la séquence cible, puis de les assembler. Ici aussi, il est nécessaire de définir une représentation simplifiée de la chaîne peptidique, de l'espace conformationnel ainsi qu'une fonction d'énergie. Depuis quelques années, les méthodes *de novo* permettent d'obtenir des modèles de bonne qualité [Aloy *et al.* 2003]. Par exemple, le groupe de D. Baker défend de bons résultats avec leur méthode *de novo* Rosetta. Cette méthode repose sur une vision du mécanisme de repliement des protéines du local vers le global : la structure locale des protéines est restreinte à un nombre limité de conformations. Des interactions non-locales stabilisent la structure globale de moindre énergie, compatible avec les conformations locales. Ils supposent que la distribution des conformations d'un fragment peut être approchée par la distribution des structures observées pour des fragments de séquence similaire dans la PDB. [Simons *et al.* 1997 ; Bystroff & Baker 1998 ; Simons *et al.* 1999a,b ; Bonneau *et al.* 2001 ; Bradley *et al.* 2003] Cette méthode a donné lieu à des développements méthodologiques comme le design de nouvelles protéines [Dantas *et al.* 2003b].

1.3.1.2 Repliement des protéines *in silico*

Le paradoxe de Levinthal [1969] dépeint une théorie de la dynamique du repliement des protéines. C. Levinthal remarqua que, en raison du très grand nombre de degrés de libertés dans une chaîne peptidique, une molécule possède un nombre astronomique de conformations possibles. De nombreuses petites protéines se replient spontanément en un temps de l'échelle d'une milliseconde voire d'une microseconde. Une protéine ne peut donc pas se replier en échantillonnant toutes les conformations possibles. En réalité, certains

chemins de repliement vers la structure finale sont nettement plus probables que d'autres. C'est ce qui rend possible la simulation du repliement *in silico*.

Les premières études *in silico* de repliement datent d'une trentaine d'années [Levitt & Warshel 1975], mais des progrès ne cessent d'être faits dans ce domaine. Comme par exemple les travaux de V. Pande qui a utilisé le calcul distribué pour simuler au delà de la microseconde de dynamique d'une petite protéine en solution. D'autres améliorations ont pu être tentées, comme l'utilisation d'une représentation implicite, et non explicite, de l'eau. Cette technique de continuum diélectrique (milieu homogène et polarisable) a permis par exemple en 2002 de prédire la structure repliée d'une petite protéine de 20 acides aminés avant la détermination expérimentale de sa structure. [Simmerling *et al.* 2002]

1.3.2 Méthodes de prédiction de structure par homologie

Devant le nombre de méthodes disponibles, nous ne pouvons toutes les détailler ici.

Les méthodes utilisant la recherche de "similarités" entre les séquences, se basent sur le principe suivant : si deux protéines partagent une forte similarité de séquence, il est très probable que cette similarité soit due à une relation d'homologie [Sander & Schneider 1991]. Ce principe permet d'inférer des connaissances sur une protéine en cherchant des homologues à celle-ci dans les bases de données de protéines connues. Deux protéines sont homologues si elles partagent un ancêtre commun. Elles peuvent avoir gardé une fonction, une structure et/ou une séquence similaires à celles de cet ancêtre. Ces méthodes se déroulent généralement en trois étapes :

1. des homologues sont recherchés dans des bases de données
2. si la structure 3D des homologues détectés est connue, un modèle 2D est construit
3. la qualité du modèle 3D est évaluée afin de proposer des raffinements.

La première étape est évidemment capitale dans la suite du processus. Plus il y a d'homologues proches trouvés, dont la structure est connue, plus la création du modèle 3D sera aisée et précise.

1.3.2.1 Les types d'alignements

La plupart des méthodes de comparaison de séquences se basent sur les techniques de programmation dynamique proposées par Needlman & Wunsch [1970], puis par Smith & Waterman [1981]. Ces techniques proposent d'aligner deux séquences de protéines en prenant les acides aminés comme éléments unitaires. Un acide aminé de la protéine A peut-être aligné face à un acide aminé de la protéine B ou face à un "trou" (*gap* en anglais). Selon la fonction de score utilisée, un score est affecté à l'alignement de deux acides aminés. De la même façon, un score est affecté à l'alignement d'un acide aminé avec un *gap*.

Nous pouvons définir trois types d'alignements (schématisés dans la figure 1.12) :

– l'alignement local

– l'alignement semi-global

– l'alignement global.

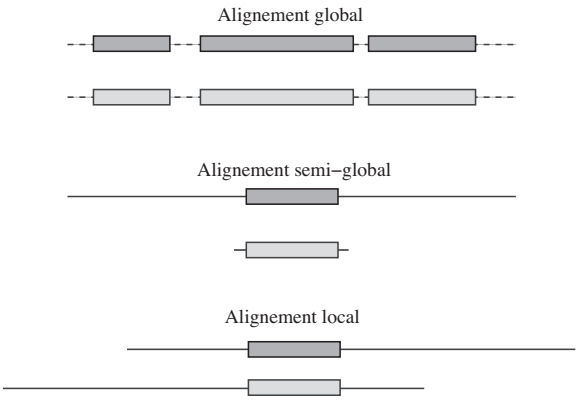


Figure 1.12 – Types d'alignement. Lors d'une comparaison de séquences de protéines, il est important de pouvoir réaliser différents types d'alignement afin de s'adapter au contexte.

1.3.2.2 Alignement local

L'alignement local est le type d'alignement le plus répandu. Il consiste à détecter des similarités locales entre deux séquences de protéines, par exemple détecter un domaine commun à deux protéines. C'est le type d'alignement le plus utilisé, en particulier dans le logiciel BLAST [Altschul *et al.* 1990] qui est basé sur une heuristique permettant de trouver des alignements locaux sub-optimaux très rapidement.

Le terme "prédiction de structure locale" englobe la prédiction de structures secondaires mais aussi d'autres types de prédiction comme la prédiction d'alphabets structuraux ou d'angles Φ et Ψ.

1.3.2.3 Alignement semi-global

De nombreuses protéines ont une structure modulaire, c'est-à-dire constituée de plusieurs modules, ou domaines. De plus, des protéines différentes peuvent partager un même domaine. Afin de détecter un domaine, autrement dit une petite séquence à l'intérieur d'une grande séquence, il convient d'utiliser un alignement semi-global. Dans ce type d'alignement, les gaps aux extrémités ne sont pas pénalisés afin de permettre au domaine de s'aligner au mieux sur la grande séquence. En revanche, les gaps à l'intérieur des séquences sont toujours pénalisés.

1.3.2.4 Alignement global

Ce type d'alignement est utilisé pour comparer des protéines d'une même famille. De telles protéines sont censées avoir des longueurs similaires. En conséquence, les gaps aux extrémités seront pénalisés.

Ces prédictions font l'objet d'une compétition internationale bisannuelle, appelée CASP (*Critical Assessment of techniques for protein Structure Prediction*) [CASP -].

Modélisation par homologie La séquence protéique cible est alignée sur toutes les séquences des protéines de la banque de données PDB à l'aide d'outils comme PSI-BLAST [Altschul *et al.* 1997]. Si deux protéines ont une identité de séquence suffisante (environ 25 % de résidus identiques ou une e-value de PSI-BLAST significative) on peut considérer qu'elles sont homologues. Alors, un modèle peut être construit sur les bases de la protéine homologue afin d'en déduire la structure 3D de la protéine cible. C'est la méthode la plus utilisée actuellement.

Le programme de modélisation par homologie le plus utilisé est MODELLER [Sali & Blundell 1993] (Pôle BioInformatique Lyonnais http ://geno3d-pbil.ibcp.fr/). À partir de l'alignement entre la séquence cible et la séquence support, MODELLER construit un modèle atomique par satisfaction de contraintes spatiales dérivées de l'alignement. La forme de ces contraintes dérive d'une analyse statistique des paires de structures homologues.

L'étape limitante de la modélisation consiste à aligner correctement les deux séquences. La modélisation des boucles variables reste difficile. Il est possible de modéliser les structures par homologie pour 16 à 30% des protéines d'un génome.

1.3.3 Évolution et homologie

Les protéines sont des adaptations biochimiques soumises à la sélection naturelle et à l'évolution. Les protéines homologues isolées à partir d'organismes différents peuvent avoir pratiquement le même repliement, mais les différences entre les séquences d'acides aminés sont d'autant plus grandes que leur distance évolutive est grande. L'évolution a donné naissance à des versions modifiées des protéines chez des organismes différents, mais elle est également à l'origine de versions modifiées d'une protéine dans un même organisme. Cela nous rappelle que les protéines sont très complexes et que nous connaissons mal la façon dont la nature transforme un message génétique linaire en une protéine tridimensionnelle. Ainsi, la recherche d'homologue en est que plus ardue.

1.3.3.1 Évolution convergente et divergente

Dans certains cas, une faible homologie de séquence combinée avec une similitude structurale élevée reflète une conservation sélective des résidus fonctionnellement importants dans des séquences ayant fortement divergé mais qui sont à l'évidence homologues malgré tout. La mandélate racémase (une enzyme de lactonisation du muconate) et l'énolase présentent une identité globale de séquence très faible, mais leurs structures et leurs sites actifs ont une forte ressemblance. Les réactions qu'elles catalysent ont en commun une étape centrale et cette étape est catalysée de la même manière, ce qui signifie qu'elles ont sans doute divergé à partir d'un ancêtre commun.

Dans d'autres cas cependant, il y a une équivalence spatiale au niveau du site fonctionnel, mais une conservation de séquence faible ou nulle des résidus fonctionnellement importants. Dans ce cas, distinguer l'évolution convergente de l'évolution divergente peut s'avérer difficile. Par exemple, les enzymes benzoylformate décarboxylase (BFD) et pyruvate décarboxylase (PDC) présentent seulement 21 % d'identité de séquence mais ont des repliements quasiment identiques. Les chaînes latérales des acides aminés catalytiques sont conservées dans leur position spatiale, au sein de la structure 2D mais pas dans la séquence. Il est possible que les deux protéines aient évolué indépendamment et aient convergé vers la même solution chimique pour résoudre le problème de la décarboxylation d'un alpha-cétoacide. Cependant, la forte similitude de leur structure globale semblerait indiquer une divergence à partir d'un ancêtre commun. Le degré d'identité de séquence est toutefois trop faible pour identifier de façon certaine la bonne possibilité.

Inversement, il existe des protéines avec des fonction biochimiques très différentes mais qui présentent des structures 3D très ressemblantes et une identité de séquence suffisante pour indiquer une homologie. Ces cas suggèrent que la structure diverge également plus lentement que la fonction au cours de l'évolution. Par exemple, la stéroïde delta isomérase, le facteur de transport nucléaire 2 et la scytalone déshydratase ont de nombreux détails structuraux communs et sont considérés comme homologues. Pourtant, l'isomérase et la

déshydratase n'ont aucun résidu catalytique essentiel en commun. Ceci suggère que ce sont les caractéristiques générales de la cavité du site actif de l'armature de ces enzymes qui ont la capacité de catalyser des réactions chimiques différentes passant par un intermédiaire énolate commun, avec des résidus distincts dans le site actif. La troisième protéine de ce groupe d'homologues (le facteur de transport nucléaire) n'a rien d'une enzyme mais sa cavité semblable à un site actif contient des résidus présents dans les sites catalytiques des deux enzymes. Par conséquent, déterminer la fonction à partir de la séquence et de la structure est compliqué par le fait que des protéines de structure similaire peuvent ne pas avoir une fonction identique, alors même qu'elles sont apparentées dans l'évolution.

1.3.3.2 Zone d'ombre

Afin de détecter une homologie, la première étape consiste à aligner une séquence de protéine de structure inconnue avec des séquences de protéines connues. Si une forte similarité est détectée, il est très probable que celle-ci soit due à une relation d'homologie [Sander & Schneider 1991]. Néanmoins, en-dessous de 30 % d'identité de séquence, les mesures de similarité de séquences ne sont plus suffisantes pour détecter des homologies [Brenner *et al.* 1998 ; Rost 1999]. Cette zone est appelée zone d'ombre ou *twilight zone* en anglais.

Les méthodes d'alignement de séquences ont montré leur efficacité [Qi *et al.* 2007]. Cependant, en dessous de 30% d'identité de séquence, des protéines homologues et non-homologues peuvent avoir des taux de similarité de séquences identiques. De plus, Brenner *et al.* [1998] montrent que, sur un jeu de 9044 paires de protéines homologues partageant moins de 40% d'identité de séquences, seules 18% sont reconnues par des méthodes d'alignement de séquences. Dans cette zone d'ombre, les méthodes d'alignement de séquence ne sont donc plus suffisamment puissantes pour détecter des homologies entre protéines [Rost 1999].

Afin de détecter des homologies dans la zone d'ombre, il convient de mettre en place des méthodes, dites de reconnaissance de repliements. Ces méthodes utilisent le fait que la structure des protéines est mieux conservée que leur séquence au cours de l'évolution [Illergard *et al.* 2009].

1.3.4 Reconnaissance de repliements

Les méthodes de reconnaissance de repliements, ou *Protein Threading* en anglais, consistent à mesurer la compatibilité entre une séquence et une structure de protéine [Bowie *et al.* 1991]. La différence avec la prédiction par homologie est que la relation d'homologie n'est pas détectable par les outils d'alignement de séquence, ce qui est le cas quand l'ancêtre commun est trop éloigné. Un autre cas peut se produire : l'évolution convergente de deux protéines vers la même structure, en l'absence d'ancêtre commun. Toutes ces méthodes comprennent les quatre éléments suivants :

- une représentation de la structure des protéines
- une fonction de score évaluant la compatibilité entre une structure et une séquence
- un algorithme d'alignement
- une statistique permettant d'évaluer la pertinence des alignements

La représentation de la structure des protéines peut prendre en compte de nombreuses informations structurelles. En effet, un acide aminé peut être enfoui au cœur de la protéine ou exposé au solvant ; être dans une hélice, un feuillet ou une boucle ; être en contact avec d'autres acides aminés ; faire partie du site actif de la protéine ; former un pont disulfure (Cystéine) ; etc. Ces informations sont directement extraites des fichiers PDB.

Des méthodes basées sur les modèles de chaînes de Markov cachées ont été proposées pour détecter l'homologie distante, tels que les programmes SAMT de Karplus *et al.* [1997, 1998] ou HMMER de Eddy [1998]. Une approche comme le logiciel FROST de Marin *et al.* [2002] consiste à superposer la séquence cible sur toutes les structures d'une banque de structures et à évaluer la compatibilité entre la séquence cible et chacune des

structures. Mais mettre en évidence la compatibilité d'une séquence avec une structure existante demeure une difficulté supplémentaire.

1.3.5 Problème inverse du repliement

C'est Eisenberg [1982] qui posera en premier le problème du repliement inverse de protéine. Il s'agit d'explorer les séquences compatibles avec un pli donné et d'identifier les plus favorables. Une façon naïve de procéder est d'effectuer des mutations au hasard sur la séquence native et de les accepter ou non selon un critère de stabilité de la protéine mutée. On se rapproche alors beaucoup du processus naturel d'évolution, qui prendrait en compte uniquement les mutations ponctuelles.

Un des enjeux actuels est de mener cette exploration de façon exhaustive pour les quelques 3 000 structures connues, ce qui nécessite une puissance de calcul importante. Des projets de plate-forme de calcul distribué comme BOINC (http ://boinc.berkeley.edu, [Dwyer *et al.* 2004 ; Kuhlman *et al.* 2003b]) participent à l'élaboration d'une cartographie complète de l'espace des séquences compatibles avec les plis connus. Ce sont des outils qui devraient se révéler puissants pour comprendre l'évolution passée des protéines et à terme, concevoir de nouvelles protéines. Déjà, cette méthode a permis, depuis 2003, d'identifier de nouvelles séquences susceptibles d'adopter un pli donné, ou d'adopter un pli tout à fait nouveau.

Chapitre 2

Prédiction de séquences théoriques par Design Computationnel de Protéines (CPD)

L'approche computationnelle permet une exploration plus exhaustive et contrôlée des mutations possibles que la génération expérimentale et aléatoire des librairies de mutants. Cette approche est connue en anglais sous le nom de *Computational Protein Design*. Nous y ferons référence par l'abréviation CPD. Par ailleurs, la notion de "design" n'a pas de correspondance exacte en français, aussi nous conserverons le terme anglais tout au long de ce manuscrit. Le CPD fut au départ essentiellement appliqué au renforcement de la stabilité de protéines existantes, le développement de protéines thermostables présentant un intérêt industriel indéniable. Par la suite, les applications du CPD furent étendues au problème inverse du repliement, au développement de nouveaux repliements ou encore de protéines présentant de nouvelles fonctions.

2.1 Introduction

L'approche historique des relations entre les séquences protéiques et leurs structures est celle du repliement, c'est-à-dire que l'on cherche à retrouver la structure (repliement

ou conformation) native d'une séquence donnée. De très nombreuses méthodes ont été développées afin de trouver des potentiels statistiques (mais également des fonctions d'énergie semi-empiriques) adaptés à ce problème. De plus, il a souvent été observé que des séquences très différentes pouvaient avoir le même repliement. Ainsi, le CPD, aussi appelé *inverse protein folding*, considère le problème dans l'autre sens : si l'on connaît un repliement, est-il possible de retrouver la séquence ou l'ensemble de séquences qui lui correspondent ?

Une protéine dispose d'un outil pour accomplir sa fonction : sa structure. Elle lui permet, par exemple, d'effectuer une réaction chimique (pour une enzyme par exemple) ou de reconnaître un ligand afin de générer une réaction appropriée de la cellule (récepteurs). Les protéines jouent un rôle majeur dans les organismes, sous la forme de récepteurs, d'enzymes, d'hormones, de régulateurs, d'anticorps... Il est possible que l'on souhaite reconstruire une protéine d'intérêt ayant un repliement particulier, par exemple dans le domaine médical. Cependant, le nombre de séquences possibles pour une structure donnée est bien trop grand pour qu'il soit possible de générer toutes les séquences admissibles pour chaque repliement d'intérêt, et les tester expérimentalement. Aussi, l'intérêt du design de protéines pourrait être ici de réduire la quantité des séquences possibles, et plus particulièrement de définir quelques séquences plus probables, à tester en priorité *in vitro*.

Le design computationnel de protéine continue à se développer comme un outil important pour la biotechnologie [Baker 2006 ; Butterfoss & Kuhlman 2006 ; Guérois & Lopez de la Paz 2007 ; Lippow & Tidor 2007 ; Pleiss 2011 ; Pantazes *et al.* 2011 ; Saven 2011 ; Samish *et al.* 2011]. De premières applications ont été menées sur des protéines apportant de nouvelles interactions avec des ligands [Looger *et al.* 2003 ; Havranek & Harbury 2003], de nouvelles activités enzymatiques [Bolon & Mayo 2001] et sur des protéines qui ont été complètement "reconçues" [Dantas *et al.* 2003a]. Au cours des dernières années, le CPD a permis la création de nouveaux plis de protéine [Kuhlman *et al.* 2003a ; Liang *et al.* 2009 ; Koga *et al.* 2012], de nouvelles enzymes [Rothlisberger *et al.* 2008 ; Jiang *et al.* 2008 ;

Richter *et al.* 2011] et l'assemblage de complexes de protéines [Saven 2010 ; Fortenberry *et al.* 2011 ; Grigoryan *et al.* 2011 ; King *et al.* 2012 ; Lanci *et al.* 2012]

2.2 Présentation du modèle

L'un des objectifs les plus courants du CPD consiste à identifier, parmi toutes les séquences possibles, celles capables de se replier dans une structure donnée. Le nombre de séquences possibles est considérable. Pour simplifier le problème, on commence par discrétiser l'espace conformationnel, en utilisant la notion de rotamères, par exemple. La procédure générale est ensuite réalisée en deux étapes. La première consiste à calculer une matrice d'énergie contenant les énergies d'interactions entre toutes les paires de résidus de la protéine en autorisant successivement tous les types d'acides aminés dans toutes leurs conformations possibles. Cette matrice d'énergie contiendra aussi l'énergie d'interaction de chaque résidu avec le squelette peptidique. La seconde étape, ou "phase d'optimisation", consiste à déterminer la combinaison optimale d'acides aminés étant donné le repliement protéique d'intérêt.

Comme nous venons de le voir, l'un des objectifs les plus courants du CPD consiste à identifier dans l'espace des séquences possibles, celles qui préserveront le repliement de la protéine d'intérêt. Or pour une petite protéine de 100 résidus en permettant les 20 acides aminés naturels à chaque position nous obtenons déjà un nombre potentiel de 20^{100} séquences. Ce nombre peut encore être augmenté si on tient compte des différentes orientations possibles des chaînes latérales et des différentes conformations de la chaîne principale. C'est pourquoi, dans le but de réduire la complexité de l'espace conformationnel, le CPD impose une discrétisation à deux niveaux :

- Discrétisation de l'espace conformationnel des chaînes latérales
- Discrétisation de l'espace conformationnel de la chaîne principale

2.2.1 Discrétisation de l'espace conformationnel des chaînes latérales

Chaque acide aminé peut être présent sous différentes conformations. La géométrie des chaînes latérales peut être définie par des angles de torsion nommés par convention χ^1 , χ^2 ,..., en partant de la chaîne principale jusqu'à l'extrémité de la chaîne latérale. Ces angles correspondent à la rotation des groupes chimiques autour des liaisons.

L'espace conformationnel peut alors être réduit en un nombre fini d'angles de torsion, adoptant une série finie de valeurs. Cette discrétisation de l'espace se prête bien aux protéines puisqu'en pratique, pour les acides aminés, certaines valeurs d'angle de torsion sont nettement plus probables que d'autres. En effet, dans des travaux pionniers, Ponder & Richards [1987] ont montré que dans les structures cristallines protéiques les chaînes latérales adoptaient un nombre limité de conformations préférentielles. Ces conformations sont aussi connues sous le nom de rotamères, concept introduit par Janin & Wodak [1978]. Chaque type d'acide aminé présente typiquement deux ou trois angles de torsions ce qui conduit à une moyenne d'environ 10 rotamères préférentiels par acide aminé. Il existe cependant une faible proportion, estimée inférieure à 5 % de conformations mal représentées par les rotamères. L'approximation rotamérique introduit donc une légère erreur dans la modélisation des protéines ; néanmoins, cette description discrète permet de réduire considérablement l'espace conformationnel, avantage essentiel pour le CPD.

2.2.2 Discrétisation de l'espace conformationnel de la chaîne principale

Dans de nombreuses études, la chaîne principale est maintenue fixée dans sa conformation native afin de simplifier l'exploration de l'espace conformationnel. La fixation du squelette peptidique a fait ses preuves dans certaines applications telles que la génération de protéines hyperstables [Malakauskas & Mayo 1998] ou le design *de novo* d'une

protéine complète [Dahiyat & Mayo 1997]. Cependant, cette approximation rencontre des limites. Par exemple, certaines chaînes latérales peuvent être considérées comme défavorables énergétiquement alors qu'un léger ajustement du squelette aurait suffit pour diminuer considérablement leur énergie.

Introduire de la flexibilité dans le squelette augmente sensiblement l'espace conformationnel. Deux approches prédominent. La première consiste à générer un ensemble de squelettes et à optimiser la séquence d'acides aminés à partir de cet ensemble de squelettes maintenus fixes durant l'optimisation. La seconde consiste à réajuster la chaîne principale pour un grand nombre de séquences fixes. Les deux approches nécessitent de spécifier à l'avance un nombre limité de conformations de la chaîne principale.

A partir de ces deux approches, différentes études furent réalisées avec plus ou moins de succès. La première fut réalisée par Crick [1953] qui s'intéressa aux protéines présentant une symétrie dans leur structure telles que les *coiled-coils* ou *TIM barrels*. La symétrie réduisant considérablement le nombre de conformations de la chaîne principale, ils purent modéliser une famille de dimères, trimères et tétramères de *right-hand coiled-coils*, repliements alors jamais observés dans la nature [Murzin *et al.* 1994 ; Harbury *et al.* 1998]. Néanmoins, cette méthode ne peut être généralisée à des repliements non symétriques. Su & Mayo [1997] traitèrent les éléments de structures secondaires comme des corps rigides capables de se déplacer les uns par rapport aux autres. Desjarlais & Handel [1999] modélisèrent explicitement la flexibilité du squelette en utilisant la combinaison d'un algorithme génétique et d'un échantillonnage Monte Carlo. Ces deux approches confirment le manque de sensibilité des fonctions d'énergie actuelles à de subtils changements de conformation du squelette.

2.2.3 Représentation de l'état déplié

Lors de son processus de repliement, une protéine va échantillonner un ensemble de conformations différentes toutes moins structurées et compactes que la native jusqu'à se

replier dans sa conformation la plus stable, appelée conformation native. Sa stabilité est mesurée par la différence d'énergie libre entre sa structure native et l'ensemble des états non-natifs. Bien que la structure d'une protéine ne soit pas figée, cette dernière adoptera sa conformation native la plus grande partie du temps. D'un point de vue thermodynamique, ce temps augmente de façon exponentielle avec l'énergie libre de repliement de la protéine. Pour un squelette peptidique donné, la séquence la plus favorable correspond à celle qui maximisera la différence d'énergie libre entre l'état replié et l'état déplié.

Modéliser l'état déplié est loin d'être évident puisque sa caractéristique reste principalement de ne pas être structuré. L'état déplié consiste en une distribution continue de conformations ou micro-états d'énergies similaires. Ces différentes conformations conduisent à l'exposition au solvant, au moins partielle, des résidus hydrophobes de la protéine. Aujourd'hui, l'état déplié demeure très difficile à caractériser par les méthodes expérimentales classiques. Certaines approches telles que le dichroïsme circulaire ou la spectroscopie infrarouge peuvent nous renseigner sur la présence d'éventuelles structures secondaires résiduelles. Toutefois, aucune méthode actuelle n'est en mesure d'apporter une information structurale pour l'état déplié avec un même niveau de détail que pour l'état natif.

Différents modèles furent proposés pour représenter l'état déplié. Un des modèles les plus simples décrit l'état déplié par une chaîne polypeptidique étendue si bien que les chaîne latérales des acides aminés interagissent essentiellement avec le solvant et les groupes voisins du squelette peptidique. En revanche, les chaînes latérales n'engagent que très peu d'interactions entre elles. Par conséquent, ce modèle considère que l'énergie libre de l'état déplié est uniquement dépendante de la composition en acides aminés et non de la séquence. L'approche la plus courante consiste à déterminer des énergies de référence pour chaque type d'acide aminé, représentant ainsi sa contribution individuelle à l'énergie libre de l'état déplié. Cette situation peut être modélisée par une collection de n tripeptides de séquence ala - X - ala où n correspond au nombre de résidus dans la protéine et

X à l'acide aminé courant d'une position donnée [Dahiyat & Mayo 1996 ; Wernisch *et al.* 2000]. C'est ce modèle de l'état déplié que nous utiliserons dans cette thèse.

Ce modèle de tripeptide pourrait conduire à omettre des structures locales induites par des interactions intra-chaînes [Ohnishi *et al.* 2004]. Pour palier ce manque, Pokala & Handel [2005] utilisèrent des fragments de 13 résidus extraits de structures natives présentant différents éléments de structure secondaire. Cependant, cette approche plus sophistiquée et plus lourde n'apporte pas d'améliorations significatives par rapport au modèle des tripeptides ; qui demeure aujourd'hui le modèle le plus communément utilisé.

2.3 Fonctions d'énergie

2.3.1 Potentiels de mécanique moléculaire classique à l'origine de ceux du CPD

Aujourd'hui, la plupart des modèles utilisés pour l'étude des protéines s'appuient sur les principes de la mécanique moléculaire. Cette description moléculaire utilise des concepts physiques simples hérités de Newton, Coulomb et Laplace. Les atomes de la protéines sont représentés par des particules sphériques avec un rayon plus ou moins incompressible et une charge nette constante dérivée de calculs de mécanique quantique ou de résultats expérimentaux. Les liaisons inter-atomiques sont considérées comme des petits ressorts avec une longueur d'équilibre équivalente aux longueurs de liaisons déterminées expérimentalement.

L'évolution d'un système de particules au cours du temps requiert une fonction d'énergie capable de décrire les forces qui guideront les différents atomes de la protéine durant la simulation et s'appuie principalement sur des potentiels physiques connus sous le nom de "champ de force". Ces potentiels sont principalement dérivés de calculs de mécanique quantique et de thermodynamique, ainsi que de données cristallographiques et spectroscopiques obtenues à partir d'un grand nombre de systèmes différents. Plusieurs années

de recherche furent nécessaires pour paramétriser ces potentiels destinés à la simulation de protéines. Ensuite, d'autres potentiels tels que les potentiels statistiques furent aussi incorporés dans les fonctions d'énergie. Ces derniers sont déduits de bases de données de structures protéiques connues. Un des avantages de ce type de fonction d'énergie est qu'il peut modéliser n'importe quel comportement déjà observé et intégré dans ces bases de données même si les mécanismes physico-chimiques de ce comportement demeurent incompris. Cependant, le grand désavantage de ces fonctions d'énergie est qu'elles sont incapables de prédire de nouveaux comportements absents des bases de données.

Certains algorithmes d'optimisation imposent des termes énergétiques décomposables en somme d'interaction de paires. C'est ainsi que dans leur forme la plus standard, les fonctions d'énergie en CPD sont constituées de termes énergétiques dits "de paires" tels que van der Waals, électrostatique, liaisons hydrogène et de termes de surface. Ainsi, l'énergie globale de la protéine sera décomposée en une somme d'énergies d'interactions de paires de résidus et d'interactions entre les différents résidus et le squelette peptidique. Ces énergies seront alors stockées dans la matrice d'énergie.

Notre fonction d'énergie prend la forme :

$$E = E_{MM} + E_{Solv} \tag{2.1}$$

$$\begin{aligned} E_{MM} \;=\; & \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{Urey-Bradley} k_u(u-u_0)^2 \\ & + \sum_{dihedrals} k_\phi\left[1+cos(n\phi-\delta)\right] + \sum_{impropers} k_\omega(\omega-\omega_0)^2 \\ & + \underbrace{\sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_{int} r_{ij}}}_{E_{Coulomb}} + \underbrace{\sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{E_{VdW}} \end{aligned} \tag{2.2}$$

$$E_{Solv} = E_{solv}^{elec} + E_{solv}^{surf} \tag{2.3}$$

Dans l'équation 2.2 nous modélisons les énergies de liaison : les déformations des liaisons E_{bonds}, des angles E_{angles} et $E_{Urey-Bradley}$, des angles dièdres $E_{dihedrals}$ et des angles impropres $E_{impropers}$. Ainsi que les énergies non liées : les énergies de van der Walls E_{vdw} et l'énergie électrostatique $E_{Coulomb}$. Ces énergies sont schématisées dans la figure 2.1.

Le terme E_{solv} sera décrit dans la modélisation du solvant implicite un peu plus tard.

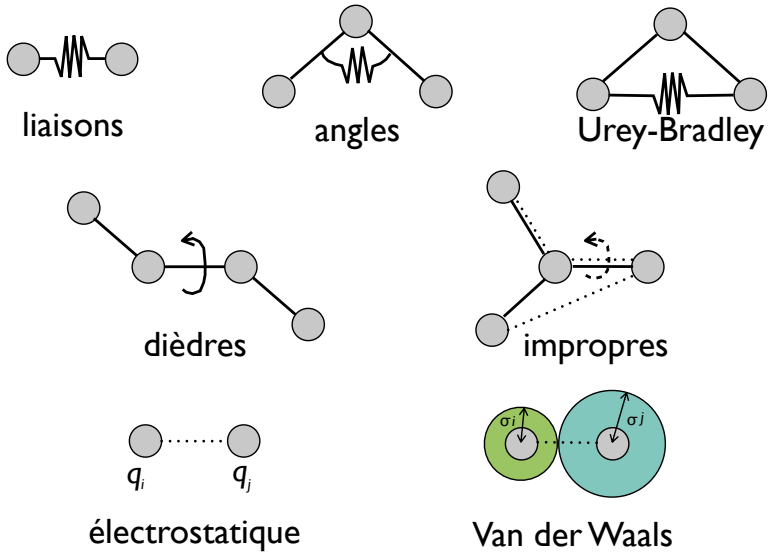


Figure 2.1 – Représentation schématique des termes contribuant à l'énergie potentielle décrite dans l'équation 2.2.

2.3.2 Différents effets influençant la stabilité de la structure

La stabilité des protéines est influencée par un certain nombre de facteurs. Les interactions qui stabilisent la structure tridimensionnelle des protéines sont principalement

des interactions faibles, non covalentes : électrostatiques, de van der Waals, liaisons hydrogène...

2.3.2.1 Énergies de van der Waals

La force de van der Waals correspond à une interaction de faible intensité entre les atomes d'une protéine. Elle est composée de deux termes : un terme attractif dominant à grande distance et un terme répulsif dominant à courte distance. Le potentiel de Lennard-Jones est une approximation mathématique utilisée couramment en modélisation pour représenter cette force. Ainsi, l'énergie de van der Waals distance-dépendante entre deux atomes a et b peut être décrite par l'équation suivante :

$$E_{VdW} = D_0 \left[\left(\frac{r_0}{r_{ab}} \right)^{12} - \left(\frac{r_0}{r_{ab}} \right)^{6} \right] \tag{2.4}$$

D_0 et r_0 sont des constantes, tandis que r_{ab} représente la distance entre les atomes a et b. Le terme répulsif en $(\frac{r_0}{r_{ab}})^{12}$ rend compte de façon *ad hoc* d'un effet purement quantique, le principe d'exclusion de Pauli qui empêche l'interpénétration mutuelle des nuages électroniques de deux atomes. En revanche, le terme attractif dérivé des interactions de dispersion, a pu être démontré rigoureusement dans le cadre de la physique quantique. Le potentiel de Lennard-Jones reste à ce jour une bonne approximation des forces de van der Waals très simple à implémenter.

Si la composante répulsive du potentiel de van der Waals est nécessaire pour éviter le recouvrement des nuages électroniques en dynamique moléculaire, elle se montre beaucoup trop restrictive dans le cas du CPD. En effet l'utilisation de rotamères discrets et d'un squelette rigide entraîne inévitablement des conflits stériques qui peuvent être évités par de petits réajustements des chaînes latérales (minimisations réalisées dans le cadre de cette thèse).

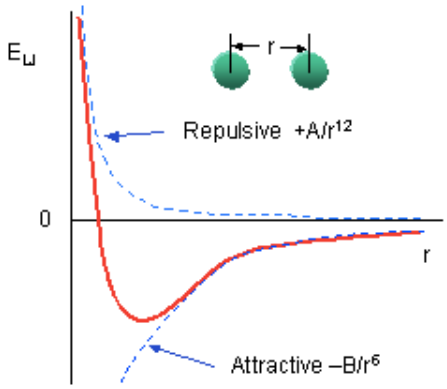


Figure 2.2 – Force de van der Waals

2.3.2.2 Liaisons hydrogène

Les liaisons hydrogène jouent un rôle important dans le maintien de la structure protéique, notamment dans les structures secondaires régulières telles que les hélices α et les feuillets β. Elles peuvent aussi avoir un rôle dans la réalisation de la fonction d'une protéine. La plupart des champs de force tels que CHARMM [Brooks *et al.* 1983], AMBER [Case *et al.* 1999] ou OPLS [Jorgensen & Tirado-Rives 1988] traitent les liaisons hydrogène de manière implicite à travers les énergies de van der Waals et électrostatiques. Elles peuvent cependant être représentées explicitement. C'est le cas du champ de force DREIDING implémenté par Dahiyat *et al.* [1997].

2.3.2.3 Électrostatique et solvatation

Les interactions électrostatiques sont d'une grande importance. Les résidus polaires ou chargés sont généralement présents à la surface des protéines et engagent un grand nombre d'interactions électrostatiques favorables avec les molécules d'eau du solvant. Toutefois, on relève la présence de quelques résidus polaires ou chargés dans le cœur des protéines. Ces derniers sont connus pour être déstabilisants du fait de leur désolvatation fortement défavorable. On peut alors supposer que ces résidus jouent un rôle important pour la

fonction et/ou le repliement final de la protéine. Ainsi, ils limitent le nombre total de conformations possibles à celles permettant des interactions suffisamment favorables (pont salins, liaisons hydrogène). *A contrario*, une protéine dont le cœur serait uniquement composé de résidus hydrophobes ne présenterait pas nécessairement un repliement unique. C'est l'observation que fit le groupe de Hill *et al.* [1990]. Lumb & Kim [1995] mirent également en évidence le rôle déterminant des résidus polaires enfouis dans l'obtention d'une structure unique.

Beaucoup d'études soulignent le rôle des interactions électrostatiques dans l'unicité du repliement et dans la réalisation de la fonction des protéines. Par conséquent, le design de nouvelles fonctions protéiques ou de nouveaux repliements nécessite des modèles physiques capables de reproduire correctement les forces électrostatiques permettant à une protéine de se replier et d'être fonctionnelle. Une protéine est le plus souvent au contact d'un milieu acqueux qui est très polarisable. Les molécules d'eau peuvent ainsi interagir directement avec des résidus hydrophiles de la protéine, entrant en compétition avec les autres groupes polaires environnants. Le solvant a alors un effet d'écrantage sur les interactions électrostatiques. Une description correcte des interactions électrostatiques implique donc la prise en compte du solvant dans la fonction d'énergie. Le modèle doit être à la fois suffisamment fiable et peu gourmand en temps de calcul. Nous détaillons les différentes représentations du solvant ci-après.

2.4 Modélisation du solvant

Le rôle de l'eau est primordial pour la structuration et la fonction des édifices biomoléculaires [Nicholls 2000]. Pour tenir compte de ses effets en modélisation moléculaire, deux grandes méthodes sont couramment utilisées. Dans la première, les molécules du solvant sont représentées de manière explicite à l'échelle microscopique, tandis que dans la seconde, seuls les effets macroscopiques du solvant sont modélisés.

2.4.1 Solvant explicite

L'utilisation de molécules d'eau explicites est certainement la meilleure et la plus réaliste des représentations du solvant disponibles actuellement dans les simulations numériques. Elle permet en effet de pouvoir réaliser des simulations stables de dynamique moléculaire à l'échelle de plusieurs dizaines de nanosecondes. De plus, la représentation au niveau atomique de ces molécules rend possible l'observation de la formation de liaisons hydrogène entre le soluté et le solvant. Ainsi, l'utilisation d'un solvant explicite est indispensable lorsque l'eau est en interaction directe avec la biomolécule et que son rôle n'est pas limité à ses effets électrostatiques.

Il existe différents modèles de molécules d'eau en modélisation moléculaire, qui diffèrent par le nombre de sites (de 3 à 5 sites pour représenter la répartition des charges dans la molécule), les différentes charges partielles de ces sites, ou la géométrie de la molécule (longueur des liaisons O-H et angle de valence $\widehat{HOH}$). Pour l'étude des biomolécules, les modèles les plus utilisés sont les modèles TIP3P [Jorgensen *et al.* 1983], SPC [Berendsen *et al.* 1981] ou SPC/E [Berendsen *et al.* 1987], qui représentent tous trois la molécule d'eau avec trois sites (l'atome d'oxygène et les deux atomes d'hydrogène).

2.4.1.1 Modèle TIP3P

Ce modèle de représentation de l'eau, que nous utilisons dans cette thèse, comprend trois atomes (un oxygène et deux hydrogènes) liés par trois liaisons (deux liaisons O-H de 0,957 Å et une pseudo-liaison H-H de 1,514 Å) de constante de force de $553\,\mathrm{kcal \cdot mol^{-1}}$. L'oxygène est chargé négativement de $-0{,}834\,\mathrm{e}$ (e est la charge élémentaire égale à $1{,}6 \times 10^{-19}\,\mathrm{C}$) et les charges des hydrogènes sont de $+0{,}417\,\mathrm{e}$. L'atténuation des interactions électrostatiques par le solvant est intrinsèque au modèle ($\epsilon = 1$) ainsi que la polarisation.

2.4.1.2 Limites du système

Le nombre de molécules d'eau à ajouter pour simuler un environnement aqueux est important et augmente avec la taille de la protéine. Il en résulte une augmentation considérable du nombre de variables du système et donc du temps de calcul. Cependant, différentes méthodes ont été employées afin de réduire le nombre d'atomes, par exemple l'usage d'un volume fini et de conditions aux limites périodiques (voir la section "Simulation de dynamique moléculaire"). Ce type de traitement du solvant sera utilisé dans nos simulations de dynamique moléculaire.

2.4.2 Solvant implicite

L'utilisation de modèles de solvant implicite est une alternative de plus en plus crédible à l'utilisation de solvants explicites. La façon la plus simple de traiter le solvant sans introduire explicitement de molécules d'eau, est de remplacer la permittivité du vide ϵ_{int} dans le terme électrostatique de l'énergie potentielle par la permittivité du milieu $\epsilon = \epsilon_{ext}\epsilon_{int}$; où ϵ_{ext} est la permittivité du milieu extérieur ou constante diélectrique. Ils permettent notamment de dépasser certaines limitations des modèles de solvant explicite en étant plus avantageux en terme de temps de calcul et en permettant l'obtention de grandeurs thermodynamiques. Deux approches différentes exploitent ces deux propriétés : le modèle Poisson-Boltzmann et le modèle de Born généralisé (*Generalized Born*).

L'effet du solvant peut être décomposé en deux : l'effet électrostatique, décrit ici par soit le modèle de Poisson-Boltzmann soit le modèle de Born généralisé, et l'effet non-polaire (coefficient surfacique).

2.4.2.1 Modèle de Poisson-Boltzmann (PB)

Actuellement considéré comme le meilleur modèle de solvant implicite, le modèle de Poisson-Boltzmann est entièrement fondé sur des concepts physiques. Les méthodes de continuum électrostatique permettent une prise en compte assez fine de l'effet du solvant.

Dans ces modèles, fondés sur l'équation de Poisson-Boltzmann ou ses approximations, la protéine est définie comme un volume de faible constante diélectrique, entourée d'un milieu infini de constante diélectrique égale à celle du solvant [Gilson & Honig 1986]. La limite entre les deux régions est déterminée par la surface moléculaire. Ce modèle repose essentiellement sur deux ingrédients physiques :

– les fortes interactions électrostatiques entre groupes chargés et solvant polarisé,

– le phénomène d'écrantage du solvant sur les interactions intra-protéine.

La polarisation induite dans le solvant est alors utilisée pour déterminer le potentiel électrostatique au sein de la protéine. La résolution de l'équation de PB est plus rapide que le traitement explicite des molécules d'eau mais reste complexe et coûteuse en terme de temps de calcul.

2.4.2.2 Modèle de Born généralisé (GB)

Le modèle de Born Généralisé [Still *et al.* 1990] reprend le même concept que celui de Poisson-Boltzmann en modélisant la protéine comme une cavité entourée d'un continuum diélectrique jouant le rôle de solvant [Bashford & Case 2000]. La contribution électrostatique à l'énergie de solvatation est alors donnée par l'équation suivante :

$$E_{solv}^{elec} = \frac{1}{8\pi}\left(\frac{1}{\epsilon_{ext}} - \frac{1}{\epsilon_{int}}\right)\sum_{i,j}\frac{q_i q_j}{f(r_{ij},a_{ij})} \tag{2.5}$$

où $f(r_{ij},a_{ij}) = \sqrt{r_{i,j}^2 + a_{ij}^2 e^{-D}}$, $a_{ij} = \sqrt{a_i a_j}$, et $D = r_{ij}^2/(2a_{ij})^2$.

Une telle forme pour la fonction f permet d'obtenir un comportement correct en $r_{ij} = 0$ et $r_{ij} = \infty$, c'est-à-dire de retrouver l'expression de Born lorsque les charges sont superposées et de tendre vers la somme des termes de Coulomb et de Born lorsque les charges sont éloignées.

ϵ est la constante diélectrique (celle de l'eau est égale à 78,5).

q_i et q_j sont les charges partielles des atomes i et j respectivement. $r_{i,j}$ est la distance entre les deux atomes i et j.

a_i et a_j sont les rayons de Born effectifs des atomes i et j respectivement. Ces rayons de Born effectifs caractérisent le degré d'enfouissement d'un atome à l'intérieur du soluté. Une estimation précise des rayons de Born effectifs est cruciale dans le modèle GB. Aussi, ils sont calculés avec la méthode de Hawkins *et al.* [1995, 1996] qui donne un rayon de Born plus grand que le rayon atomique.

Tsui & Case [2000] ont montré que pour les protéines, cette modélisation du solvant permettait un gain important de temps par rapport à l'utilisation d'un solvant explicite tout en représentant raisonnablement des effets de solvant [Xia *et al.* 2002]. L'avantage de ce modèle continu est de pouvoir limiter le nombre d'atomes du système par rapport à l'utilisation d'un solvant explicite tout en tenant compte des effets électrostatiques du solvant. Pour une protéine d'environ 2000 atomes, le remplacement du solvant explicite par l'approche de Born Généralisée représente ainsi un gain d'environ 30 % de temps de calcul (sans autres simplifications).

Archontis & Simonson [2005] approximèrent le modèle classique de GB par une approche originale. Les résidus sont considérés comme des unités à part entière puisqu'on ne définit plus des rayons de solvatation atomique mais de résidus (B_i). L'énergie "propre" de chaque résidu est décomposable en terme d'énergie de paires de résidus ce qui n'est pas le cas du terme énergétique qui décrit l'effet d'écrantage du solvant. Ce terme dépend de la configuration globale de la protéine. Cette méthode apporte des résultats comparables au modèle de GB classique et se révèle très encourageante pour les problèmes de CPD.

2.4.2.3 Effets non polaires

Lorsque l'effet électrostatique de solvatation est traité par le simple terme de Coulomb, le modèle de solvant implicite est appelé CASA (*Coulombic Accessible Surface Area*). Lorsqu'il est traité par un terme GB, le modèle est noté GASA. Nous rappelons que

l'énergie $E = E_{MM} + E_{solv}$ et que $E_{Solv} = E_{solv}^{elec} + E_{solv}^{surf}$ où le terme surfacique $E_{solv}^{surf} = \alpha \sum_i \sigma_i A_i$.

Modèle CASA Sa simplicité et son efficacité font de lui le modèle le plus répandu dans le domaine du CPD. Initialement implémenté par Wesson et Einsenberg, ce modèle utilise une constante diélectrique afin de mimer l'écrantage des interactions électrostatiques protéine-protéine due à la présence du solvant très fortement polarisé. A ce terme électrostatique, il faut ajouter un terme dépendant de la surface accessible au solvant. Les différents types atomiques sont alors caractérisés par des paramètres de solvatation atomique dérivés de calculs expérimentaux d'énergie libre de transfert octanoleau ou vapeur-eau. Ces paramètres vont refléter le caractère hydrophile/hydrophobe de chaque type atomique. La contribution énergétique de chaque atome est donnée par le produit de son paramètre de solvatation atomique et de sa surface accessible au solvant. Ce terme va favoriser l'exposition des groupes polaires.

Nous rappelons que E_{MM} contient le terme de Coulomb, par conséquent le modèle CASA peut alors être représenté par E_{MM} et le terme surfacique :

$$E_{CASA} = E_{MM} + \underbrace{\alpha \sum_i \sigma_i A_i}_{E_{solv}^{surf}} \tag{2.6}$$

Modèle GASA Puisque E_{MM} contient le terme de Coulomb, il faut définir le modèle GASA par l'équation :

$$E_{GASA} = E_{MM} + \underbrace{\frac{1}{8\pi}\left(\frac{1}{\varepsilon_{ext}} - \frac{1}{\varepsilon_{int}}\right) \sum_{i<j} \frac{q_i q_j}{f(r_{ij}, a_{ij})}}_{E_{solv}^{elec}} + \underbrace{\alpha \sum_i \sigma_i A_i}_{E_{solv}^{surf}} \tag{2.7}$$

ε correspond à la constante diélectrique du milieu pour réduire les interactions électrostatiques. Le terme de droite correspond à la somme des surfaces accessibles au solvant

A_i de tous les atomes i de la protéine, pondérées par leurs paramètres de solvatation atomique respectifs σ_i ($\text{kcal} \cdot \text{mol}^{-1} \cdot \text{Å}^{-2}$). Ce terme est lui même pondéré par le facteur surfacique α, qui pour l'ensemble de nos calculs, aura une valeur de 1. Les surfaces sont calculées par l'algorithme de Lee & Richards [1971].

2.4.3 Algorithmes d'optimisation appliqués au CPD

Le CPD nécessite des algorithmes d'optimisation capables de déterminer les meilleures solutions parmi le très grand ensemble des séquences et des conformations possibles. Ces algorithmes doivent faire un compromis entre la rapidité d'exécution et l'exhaustivité de l'exploration. En outre, le choix de la méthode est fortement dépendant de la représentation de l'espace conformationnel et de la fonction d'énergie employée.

Il existe deux types d'algorithmes d'optimisation. Les algorithmes déterministes, aussi connus sous le nom d'algorithmes "semi-exhaustifs", convergent systématiquement vers une unique solution pour un jeu de paramètres donné. Ces méthodes requièrent toutes une représentation discrète de la chaîne principale et des chaînes latérales, et sont restreintes à des fonctions d'énergie décomposables en une somme d'énergies de paires. Les algorithmes stochastiques ou semi-aléatoires échantillonnent de façon aléatoire l'ensemble des solutions possibles.

2.4.3.1 Méthodes déterministes ou semi-exhaustives

Les méthodes réellement exhaustives sont restreintes à de petits espaces conformationnels. La complexité combinatoire du CPD impose donc des méthodes semi-exhaustives qui autorisent uniquement certaines conformations discrètes afin de réduire l'espace conformationnel. Deux algorithmes prédominent dans le domaine du CPD : le champ moyen et le Dead-End Elimination (DEE). Par souci de concision, nous détaillerons seulement l'exemple du champ moyen, car utilisé dans notre équipe pour d'autres applications.

Un exemple : le champ moyen. Le champ moyen fut au départ appliqué à la prédiction de l'orientation des chaînes latérales mais a ensuite vu son champ d'action s'étendre à l'ambitieux problème du CPD. L'idée principale du champ moyen est de "résumer" toutes les interactions possibles entre un rotamère donné et tous les autres rotamères de la protéine en une unique interaction moyenne. Cette méthode utilise donc une représentation où chaque rotamère interagit avec toutes les conformations possibles des chaînes latérales environnantes, pondérées par leur probabilité respective.

Le champ moyen calcule itérativement la probabilité de Boltzmann $P(i,k)$ de chaque rotamère k pour chaque position i à partir de son énergie $E(i,k)$:

$$P(i,k) = \frac{e^{-\frac{E(i,k)}{RT}}}{\sum_{l=1}^{N_i} e^{-\frac{E(i,l)}{RT}}} \tag{2.8}$$

R est la constante des gaz parfaits tandis que T est la température, et N_i le nombre de rotamères à la position i. L'énergie $E(i,k)$ a la forme :

$$E(i,k) = E_{BB}(i,k) + \sum_{j \neq i} \sum_{l} E(ik,jl) P(j,l) \tag{2.9}$$

Le premier terme E_{BB} représente l'énergie d'interaction avec le squelette. Le second terme décrit les énergies d'interaction entre le rotamère (i,k) et tous les rotamères l pour chaque position j de la chaîne ponderées par leur probabilité $P(j,l)$. $E(i,k)$ correspond donc à la moyenne des énergies d'interaction entre la chaîne latérale (i,k) et son environnement. Au départ, tous les rotamères d'une position donnée peuvent être considérés comme équiprobables. On calcule ensuite les énergies de chaque rotamère en fonction des probabilités des chaînes latérales avoisinantes. De nouvelles probabilités de Boltzmann pour chaque rotamère (i,k) sont alors déduites de ces énergies $E(i,k)$.

Ce processus est répété jusqu'à convergence des probabilités et des énergies. Afin d'éviter un phénomène d'oscillations, la convergence peut être considérablement améliorée en tenant compte d'une "mémoire" du cycle précédent lors de la mise à jour des probabilités. Cette nouvelle condition conduit à l'expression suivante :

$$P(i,k)^{(n+1)} = \lambda P(i,k)^{(n)} + (1-\lambda)P(i,k)^{(n-1)} \tag{2.10}$$

Le champ moyen ne garantit pas la convergence vers le minimum d'énergie globale mais vers un ensemble de rotamères, où chaque rotamère est le plus probable pour une position donnée. L'avantage de cet algorithme est son temps d'exécution relativement rapide. Ce temps augmente linéairement avec le nombre de résidus de la protéine, permettant l'application du champ moyen a des systèmes de grande taille.

2.4.3.2 Méthodes stochastiques ou semi-aléatoires

Les méthodes stochastiques, échantillonnent aléatoirement l'espace des séquences et des structures en se déplaçant d'une solution à l'autre d'une façon dépendante du paysage énergétique et des lois imposées par l'algorithme d'optimisation. Plus faciles à implémenter, les méthodes stochastiques sont aussi beaucoup plus rapides que les méthodes exhaustives. Le principal algorithme utilisé en CPD est l'algorithme de Monte Carlo.

Un exemple : Monte Carlo Le Monte Carlo est l'une des méthodes stochastiques les plus simples à implémenter. Son principe est de proposer itérativement une modification au modèle étudié puis de décider d'accepter ou rejeter cette modification selon le critère de Metropolis. Cette méthode peut être utilisée pour optimiser des séquences d'acides aminés, des orientations de chaînes latérales, des conformations de chaînes principales ou encore tous ces critères simultanément.

Dans le cas du CPD, un acide aminé dans un rotamère donné est aléatoirement modifié pour une position choisie au hasard dans toute la séquence. La nouvelle énergie du sys-

tème E_{new} est alors mise à jour. Si cette énergie est plus faible que l'énergie E_{old} de l'état précédent alors la modification est acceptée. Si au contraire cette énergie se trouve plus élevée, la perturbation est acceptée avec la probabilité $p = \exp^{\frac{E_{new}-E_{old}}{RT}}$. Cette opération est répétée un grand nombre de fois. La méthode de Monte Carlo permet de franchir les barrières d'énergie et ainsi de surmonter les multiples minima locaux du paysage énergétique. La température peut être ajustée pour faciliter le franchissement de barrières d'énergie. Le recuit simulé reprend ce principe en chauffant le système puis en le refroidissant afin de diminuer graduellement la probabilité d'accepter des conformations de haute énergie.

2.4.4 Simulation de dynamique moléculaire

La dynamique moléculaire est une technique qui s'est considérablement enrichie ces trois dernières décennies. Initialement limitée aux systèmes atomiques et moléculaire, elle permet de simuler des systèmes très divers, en particulier, des molécules d'intérêt biologique de grande taille. La dynamique moléculaire est la méthode la plus intuitive et la plus naturelle pour explorer la surface d'énergie potentielle d'une biomolécule. Historiquement, la première molécule d'intérêt biologique (l'inhibiteur de la trypsine pancréatique bovine BPTI) a été modélisée par une simulation de dynamique moléculaire de 9,2 ps, il y moins de 30 ans [McCammon *et al.* 1977] et cette technique de simulation est toujours largement utilisée actuellement [Karplus & McCammon 2002]. La dynamique moléculaire cherche à obtenir une exploration de la surface d'énergie potentielle, l'accumulation de statistiques et bien évidemment, la construction d'une dynamique réelle, par exemple de repliement [Karplus & McCammon 2002].

Pour de nombreux systèmes atomiques, les temps de relaxation des phénomènes sont très inférieurs à 10×10^{-8} s et la dynamique moléculaire est un très bon outil. Les échelles de temps rencontrées dans les systèmes biologiques peuvent couvrir jusqu'à 16 ordres de grandeurs (tableau 2.1) [Leach 2001 ; Schlich 2002]. Le défi des simulations de biomolécules

Mouvement	Temps (s)
Elongation d'une liaison	10^{-14}
Flexion d'une liaison	10^{-14}
Rotation des chaînes latérales de surface	$10^{-11} - 10^{-10}$
Rotation des chaînes latérales intérieures	$10^{-14} - 1$
Repliement de protéines	$10^{-6} - 10^{2}$

Table 2.1 – Principales échelles de temps rencontrées en dynamique moléculaire pour des systèmes biologiques.

est de modéliser des phénomènes lents comme le repliement d'une protéine tout en tenant compte de mouvements rapides comme la vibration d'une liaison. Plusieurs approches tentent de résoudre ce problème. Les algorithmes de dynamique avec contraintes, SHAKE [Ryckaert *et al.* 1977] et RATTLE [Andersen 1983], modifient les équations du mouvement de façon à geler les vibrations les plus rapides. Dans la méthode à pas de temps multiples [Streett *et al.* 1978], les forces à longue distance sont évaluées moins souvent que celles à courte distance. Avec ces stratégies et une augmentation significative de la puissance de calcul, la dynamique moléculaire a pu plus récemment simuler des biomolécules sur des temps de l'ordre de centaines de nanosecondes ou des systèmes allant jusqu'au million d'atomes [Karplus & McCammon 2002]. Ces performances sont remarquables mais restent cependant insuffisantes pour simuler le repliement d'une protéine qui peut se dérouler en plusieurs dizaines de microsecondes.

Plus précisément, la dynamique moléculaire consiste à étudier la trajectoire d'une molécule en appliquant les lois de la mécanique classique newtonienne, c'est à dire à simuler les mouvements atomiques au cours du temps. Ces mouvements correspondent à des vibrations autour d'un minimum ou au passage d'un minimum à un autre minimum d'énergie. Ainsi la dynamique moléculaire permet de s'extraire d'un minimum local.

Comme les générations de séquences par CPD se déroulent avec des structures assez fortement contraintes (pour rappel, le squelette peptidique est fixe) ainsi qu'une définition implicite du solvant, nous complétons alors nos prédictions structurales par des simulations de dynamique moléculaire, avec un solvant défini explicitement.

Les simulations de dynamique moléculaire se déroule principalement en deux étapes : la préparation du système (avec l'ajout d'un environnement aqueux et des hydrogènes manquants) puis l'exploration conformationnelle. Entre les deux phases, une minimisation est généralement effectuée pour relaxer le système, suivie d'une étape de chauffage pour atteindre la température de simulation souhaitée, et enfin une équilibration permet au système de se stabiliser avant de commencer l'exploration de conformations. Les analyses sont réalisées à partir de l'exploration uniquement.

L'algorithme général Les conformations du système au cours du temps sont obtenues en intégrant les équations de Newton du mouvement. La trajectoire est obtenue en résolvant l'équation différentielle issue de la deuxième loi de Newton $F = ma$ qui relie l'accélération à la force :

$$\frac{\delta^2 x_i}{\delta t^2} = -\frac{1}{m_i}\frac{\delta E}{\delta x_i} \tag{2.11}$$

Cette équation décrit le mouvement d'un atome i, de masse m_i, selon un degré de liberté x_i, et une énergie potentielle E (décrite en 2.2).

Différents algorithmes permettent de réaliser l'intégration numérique des équations de Newton en utilisant la méthode des différences finies. L'idée principale est de séparer l'intégration en étapes élémentaires, chacune étant espacée par un pas de temps fixe δt. La force, considérée constante durant le pas d'intégration, exercée sur chaque particule au temps t est calculée comme la somme vectorielle des interactions avec les autres particules. La force est alors utilisée pour déterminer les accélérations, qui sont combinées aux vitesses et aux positions au temps t pour calculer les positions et vitesses au temps $t + \delta t$. Les forces sur les particules en leurs nouvelles positions sont alors calculées et le processus est itéré.

Tous les algorithmes d'intégration considèrent que les positions, vitesses et accélérations peuvent être approchées par un développement de Taylor-Young. Le plus utilisé est l'algorithme de Verlet [1967]. Il utilise les positions et accélérations au temps t et les positions à l'étape précédente $r(t-\delta t)$, pour calculer les nouvelles positions $r(t+\delta t)$. Les relations utilisées sont issues d'un développement au second ordre :

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{2.12}$$

$$r(t-\delta t) = r(t) - \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{2.13}$$

On obtient alors en additionnant ces deux équations :

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2 a(t) \tag{2.14}$$

L'algorithme de Verlet est simple à mettre en œuvre mais présente plusieurs inconvénients. En particulier, l'absence d'un terme explicite pour les vitesses nécessite de les calculer séparément. De plus, des problèmes de précisions sont rencontrés en raison de l'ajout d'un terme du second ordre $\delta t^2 a(t)$ à deux termes du premier ordre beaucoup plus grands.

Une des variantes de l'algorithme de Verlet est aussi souvent utilisé, il s'agit de l'algorithme de Leap-frog [Hockney & Eastwood 1988]. Dans cet algorithme, les vitesses au temps $t + \frac{1}{2}\delta t$ sont utilisées pour calculer les positions au temps $t + \delta t$. Les vitesses et les positions ne sont pas déterminées au même moment, les unes étant décalées de $\frac{1}{2}\delta t$ par rapport aux autres. Cette définition a donné le nom de "saute-mouton" à l'algorithme.

D'autres variantes de l'algorithme de Verlet existent, comme celle appelée Velocity-Verlet [Swope *et al.* 1982].

Le choix du pas d'intégration δt est très important pour mener à bien une dynamique moléculaire. Il est choisi le plus long possible afin de limiter le temps de calcul mais doit rester faible par rapport à l'échelle de temps des mouvements des atomes au cours de la dynamique. Un pas d'intégration d'une femtoseconde est généralement considéré comme acceptable.

Conditions aux limites La plupart des phénomènes étudiés ont lieu en solution et le traitement du solvant est un point critique des simulations de dynamique moléculaire. La méthode des conditions aux limites périodiques permet d'employer une solvatation explicite tout en gardant un nombre fini de molécules. La protéine est alors plongée dans une boîte d'eau, qui est reproduite à l'identique dans toutes les directions de l'espace. Lorsqu'un particule sort de la boîte centrale par une des faces, elle est réintroduite par la face opposée. Cette méthode permet donc de simuler un système plongé dans un milieu de solvant infini.

2.5 Quelques applications du CPD

Comprendre ou prédire les structures et interactions moléculaires est un premier pas vers leur ingénierie. Générer des protéines en augmentant leur stabilité est une des applications du CPD. D'autres applications tentent d'améliorer la catalyse de certaines enzymes existantes, de modifier ou de générer des affinités spécifiques pour des ligands, des substrats, des peptides ou d'autres protéines, ou encore espèrent générer de nouvelles protéines et enzymes.

Le CPD devient désormais un outil puissant et son efficacité est souvent démontrée en ce qui concerne la génération de protéines améliorées, voire nouvelles. Ainsi nous acquérons une meilleure compréhension des protéines et de leurs fonctions à travers de nombreuses applications.

2.5.1 Étude d'interactions biologiques

2.5.1.1 Interactions protéine-protéine

La plupart des protéines s'assemblent en complexe pour accomplir leur fonction. Cela illustre l'importance des interactions entre protéines dans une cellule. Actuellement, les principes qui régissent ces interactions ne sont que partiellement compris, assez mal décrits et donc mal prédits. Les prédictions peuvent cependant fonctionner comme le montre l'exemple de l'interaction entre le domaine TEM-1 de la *β-lactamase* avec le *β-lactamase inhibitory protein* BLIP [Strynadka *et al.* 1996].

La prédiction des interactions entre protéines nécessite l'utilisation de programmes capables de prédire la configuration des complexes biologiques. Un certain nombre de travaux ont été réalisés à partir de l'étude des complexes cristallographiques présents dans la PDB. [Carugo & Argos 1997 ; Dasgupta *et al.* 1997 ; Janin & Rodier 1995 ; Jones & Thornton 1996 ; Vajda *et al.* 2002]. Malheureusement, ils comportent des complexes naturels et des complexes artificiels engendrés par les conditions de cristallisation. La distinction des deux familles n'est pas triviale, et pourtant, indispensable à la réalisation de prédictions fines. La notion de prédiction implique l'analyse de données rassemblées, de préférence, en bases de données de complexes protéiques observés expérimentalement. Ces bases sont essentielles pour étudier les interfaces protéiques et développer des programmes de prédiction. Gray et Mendez ont établi des revues plus exhaustives des programmes et méthodes actuels [Gray 2006 ; Mendez *et al.* 2005].

Des sessions équivalentes à CASP ont été créées en 2001 pour évaluer les méthodes de prédiction des interactions entre protéines (session CAPRI). Elles se distinguent par leur faible nombre de cibles et de participants comparé aux CASPs. La prédiction des interactions entre protéines reste du domaine d'expertise académique et est bien loin d'être arrivée à maturité comme l'est la prédiction de structures 3D par homologie ou

bien la prédiction des interactions protéine-petite molécule. L'objectif de ces sessions est de faire progresser ce domaine de recherche comme cela a été le cas pour les CASPs.

Il existe d'autres cas particuliers à l'étude comme les complexes résultants de la formation d'un pont disulfure ou bien les complexes dont la zone d'interface inclut également un ligand, de l'ADN ou de l'ARN.

2.5.1.2 Interactions protéine-ligand

L'équipe de Simonson *et al.* effectue des mutations aléatoires dans la structure de la protéine concernée, qu'on accepte ou qu'on rejette selon leur effet sur la stabilité et sur la reconnaissance protéine-ligand. Ils s'efforcent de modifier la spécificité de plusieurs enzymes pour leurs substrats naturels. Récemment, ils ont obtenu des variantes de la tyrosylARNt synthétase ayant une activité détectable pour des substrats autres que le substrat naturel [Lopes *et al.* 2009]. Cette « validation de principe » devrait maintenant ouvrir un large champ d'applications en biotechnologie et biologie synthétique.

2.5.2 Design de protéines entières

Actuellement, on compte peu d'exemples de *design* de protéines entières. Cusack *et al.* [1990] ; Dahiyat & Mayo [1997] furent les premiers à relever ce défi, avec l'ingénierie d'un motif $\beta\beta\alpha$ de 28 résidus structurés en doigts de zinc. Aucune contrainte sur la séquence n'avait été ajoutée ce qui était novateur pour l'époque. La séquence obtenue présentait un score d'identité de 21 % avec la séquence native. Les auteurs constatèrent sans surprise que 75 % de ces identités étaient localisées dans le cœur, les résidus de surfaces se révélant beaucoup moins bien conservés. Malgré ce faible score d'identité, leur séquence fut capable de se replier *in silico* dans une structure compacte et stable très similaire à la structure native. Après ce succès, d'autres prouesses similaires se succédèrent telles que le design complet de différents variants d'homéodomaines [Marshall & Mayo 2001] ou de la séquence complète du domaine WW [Kraemer-Pecore *et al.* 2003]. Cependant ce type d'étude n'est

pas très répandu. En effet, selon les applications, redessiner la protéine dans sa totalité peut se révéler utile ou au contraire infructueux et coûteux. On note toutefois quelques applications telle que la reconnaissance de pli nécessitant le design de la protéine dans son intégralité.

Travaux de D. Baker (Rosetta) Le groupe de D. Baker réalise de nombreux succès dans ce domaine. Ils développèrent une méthode qui explore alternativement l'espace des séquences et des structures : ils optimisent d'abord les séquences pour un squelette peptidique donné fixe, puis optimise le repliement du squelette étant donnée une séquence d'acides aminés [Kuhlman *et al.* 2003b ; Saunders & Baker 2005].

L'une des hypothèses fondamentales à la base de la méthode Rosetta est que l'on peut obtenir une approximation raisonnable de la distribution des conformations possibles pour un court segment donné de séquence, à partir de la distribution des structures adoptées par la séquence et par des séquences étroitement apparentées dans des structures protéines connues. Les banques de fragments pour de courts segments de la chaîne sont construites à partir des bases de données de structures de protéines. A aucun moment on n'utilise la structure native globale pour sélectionner des fragments ou identifier des segments de la structure. On recherche ensuite l'espace conformationnel défini par ces fragments à l'aide de la procédure de Monte Carlo utilisant une fonction énergétique qui favorise les structures compactes avec des brins appariés et des résidus hydrophobes enfouis. Au total, 1000 simulations indépendantes sont réalisées pour chaque séquence requête (*query* en anglais) et les structures résultantes sont regroupées. On a simplement choisi les centres des plus gros groupements comme modèles ayant le degré de confiance le plus élevé. Les centres de ces groupements ont ensuite été classés en fonction de la taille des groupements qu'ils représentent, avec les centres des groupements les plus importants désignés comme modèles avec la confiance la plus élevée. Avant le regroupement, la plupart des structures produites par Rosetta sont incorrectes (c'est-à-dire que les structures justes

représentent moins de 10 % des conformations produites). Pour cette raison, la plupart des conformations créées par Rosetta sont qualifiées de formes plausibles (figure 2.3). Le problème de la discrimination entre structures justes et structures plausibles mais erronées dans les population de Rosetta est toujours à l'étude. Pourtant, certains tests de calcul ont montré que le meilleur centre de groupement s'accorde bien avec le repliement global de la protéine.

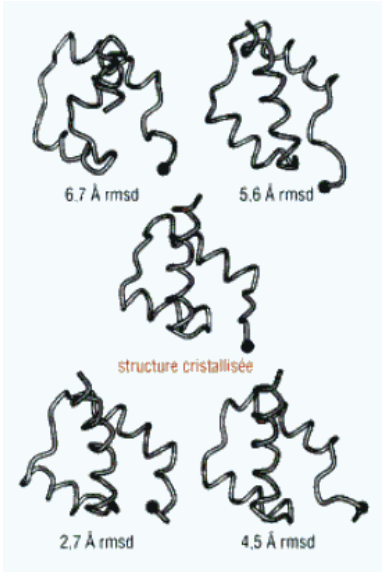


Figure 2.3 – Certaines structures plausibles produites par la méthode Rosetta. La structure d'un homéodomaine déterminée expérimentalement est au centre. Les autres structures sont produites par l'approche de Monte Carlo dans Rosetta, en utilisant uniquement la séquence de la protéine. Même si certaines structures sont très éloignées de la véritable structure, les autres sont suffisamment approchantes pour que l'on reconnaisse le motif de repliement [Simons *et al.* 1997].

Ainsi, cette méthode a permis à Kuhlman *et al.* [2003b] de créer un nouveau repliement protéique absent de la PDB, ou plus récemment l'équipe de Baker [Richter *et al.* 2011].

Chapitre 3

Sujet d'étude : les domaines SH3

La définition des domaines d'homologie date des premières études sur les protéines à activité tyrosine kinase (PTK). Les domaines les plus répandus et les plus étudiés sont les domaines SH2 et SH3 (*Src Homology* 2 et 3), souvent trouvés de concert, voire en plusieurs exemplaires dans une même protéine. De nombreux articles et revues ont été consacrés aux domaines SH2 et SH3 [Saksela & Permi 2012]. Comme les domaines SH2, les domaines SH3 n'ont pas de site topologique fixe dans leurs protéines hôtes, en accord avec leur caractère de module discret. Ils se retrouvent souvent dans les organismes vivants ; presque tous ceux que l'on connaît chez les mammifères sont aussi présents chez le nématode et la drosophile.

De part sa petite taille (une soixantaine de résidus seulement) et son repliement caractéristique en feuillets β, le domaine d'homologie 3 Src (domaine SH3) paraissait un bon candidat pour étudier une famille de structures et en faire un modèle d'étude pour la génération de séquences théoriques. Le rôle essentiel des domaines SH3 est de permettre une reconnaissance spécifique avec d'autres protéines et la transmission d'un message en réponse à un stimulus. Les principales interactions impliquées par les domaines SH3 concernent la signalisation cellulaire, mécanisme que nous allons brièvement décrire ici.

3.1 Signalisation intra-cellulaire et domaines SH3

3.1.1 Notion de domaine

Comme nous l'avons déjà mentionné (chapitre 1), les domaines sont des régions protéiques, que l'on définit suivant plusieurs caractéristiques qui se superposent parfaitement :

- sur le plan de l'évolution des protéines, les domaines sont les éléments conservés
- sur le plan fonctionnel, les domaines sont les sous-parties assurant les différentes activités élémentaires des protéines, l'activité finale d'une protéine résultant de la combinaison des activités apportées par les différents domaines la composant,
- sur le plan structural, les domaines sont, au sein des protéines, séparés les uns des autres, et résultent chacun du repliement complexe de la chaîne polypeptidique. En revanche, les domaines sont indépendants structuralement les uns des autres, et un fragment issu d'un domaine n'intervient pas dans la structure secondaire d'un autre domaine. Cependant, les différents domaines d'une même protéine peuvent interagir entre eux et moduler leurs activités respectives en fonction d'événements cellulaires,
- sur le plan de l'étude des protéines, il est généralement possible de construire des chimères en réalisant des protéines de fusion composées de domaines de différentes protéines, et d'expliquer les résultats de ces combinaisons nouvelles de fonctions élémentaires. Il est également très courant et justifié d'étudier structuralement les différents domaines d'une même protéine indépendamment les uns des autres.

Il a rapidement été remarqué une similarité de la fonction des protéines composées de groupes de domaines identiques. Ce qui pouvait être une hypothèse assez naturelle, de par la similarité implicite des séquences primaires, confirme l'existence d'une coopération fonctionnelle entre les domaines. Gerstein & Hegyi [2001] ont étudié la similarité fonctionnelle de protéines partageant les mêmes domaines SCOP dans différentes espèces eucaryotes. Ils montrent que :

- deux tiers des protéines monodomaines composées du même domaine ont une fonction similaire,
- 35% des protéines multidomaines possédant un domaine similaire, ont des fonctions semblables,
- ce taux monte à 80% si les protéines multidomaines ont deux domaines distincts en commun (sans tenir compte de l'ordre séquentiel de ces domaines),
- si elles ont une composition strictement identique (ordre et nombre d'occurences identiques), 90% des protéines multidomaines ont la même fonction. Les 10% restants peuvent s'expliquer par le fait qu'il existe différentes configurations spatiales pour une même séquence de domaine.

Ces observations sont confirmées par Ye & Godzik [2004] sur les trois domaines du Vivant. A l'aide de réseaux de domaines co-occurents, ils forment des classes de groupes de domaines récurrents et constatent que les protéines appartenant à une même classe, ont tendance à avoir des fonctions similaires.

Il existe un grand nombre de types de domaines protéiques différents. La figure 3.1 présente les principaux domaines d'interaction protéine-protéine impliqués dans la signalisation cellulaire : SH2, PTB, SH3, WW, PH, BH3 et EH. Le sujet de cette thèse concernant tout particulièrement l'étude des domaines SH3, nous les décrirons donc spécifiquement dans ce chapitre.

Les domaines SH3 (pour *Src Homology Domain type 3*) sont de petits domaines protéiques d'environ 50 à 75 acides aminés, très structurés et compacts. Ils présentent de fortes homologies de structure entre eux. La conception d'agents thérapeutiques ciblant ces domaines a fait l'objet de nombreuses publications. Dalgarno *et al.* [1997] ; Vidal *et al.* [2001]

3.1.2 La signalisation intra-cellulaire

La signalisation cellulaire est l'un des mécanismes fondamentaux dirigeant la vie, la mort et bien sûr la division des cellules. Il existe une très grande diversité de modalités de communication intra-cellulaire. Néanmoins, nous nous focalisons uniquement sur les signaux transmis par des protéines, et nous ne parlerons pas des signaux transmis par des acides nucléiques ou d'autres molécules (ions, lipides, ...). La première et principale distinction que l'on peut faire entre les modes de signalisation impliquant des protéines est la suivante :

- Signalisation **par action enzymatique** : le signal est transmis d'une protéine A à une protéine B par modification de B sous l'action d'une région de A. (modifications post-traductionnelles, phosphorylation/déphosphorylation, clivage protéolytique, modifications de chaînes latérales, oxydation/réduction ...)
- Signalisation **sans action enzymatique** : le seul contact entre les protéines A et B suffit à la transmission du signal, sans que A n'entraîne de modification durable de B. (modification temporaire de la conformation de B pour par exemple activer ou modifier son état d'activation catalytique)

Un des modes les plus répandus au sein de la voie de signalisation impliquant les protéines, et celui qui nous intéresse plus particulièrement, est l'interaction protéine-protéine.

Tous les domaines présentés dans la figure 3.1 participent à la transmission de signaux intracellulaires, sans modifier les protéines auxquelles il se fixent. Ces domaines sont relativement fréquents dans les protéines de signalisation, comme on peut le voir dans la figure 3.2. Ces protéines sont constituées de modules ou domaines dont certains sont représentés sur ce schéma. On peut noter que ces interactions se font presque toujours de façon asymétrique : par exemple lors d'une interaction entre le domaine SH2 d'une protéine A et une seconde protéine B, la région de B impliquée dans l'interaction n'est pas nécessairement et probablement jamais un domaine SH2. En revanche, il peut s'agir d'une petite partie d'une domaine quelconque de B, et dont les propriétés vont être profondément

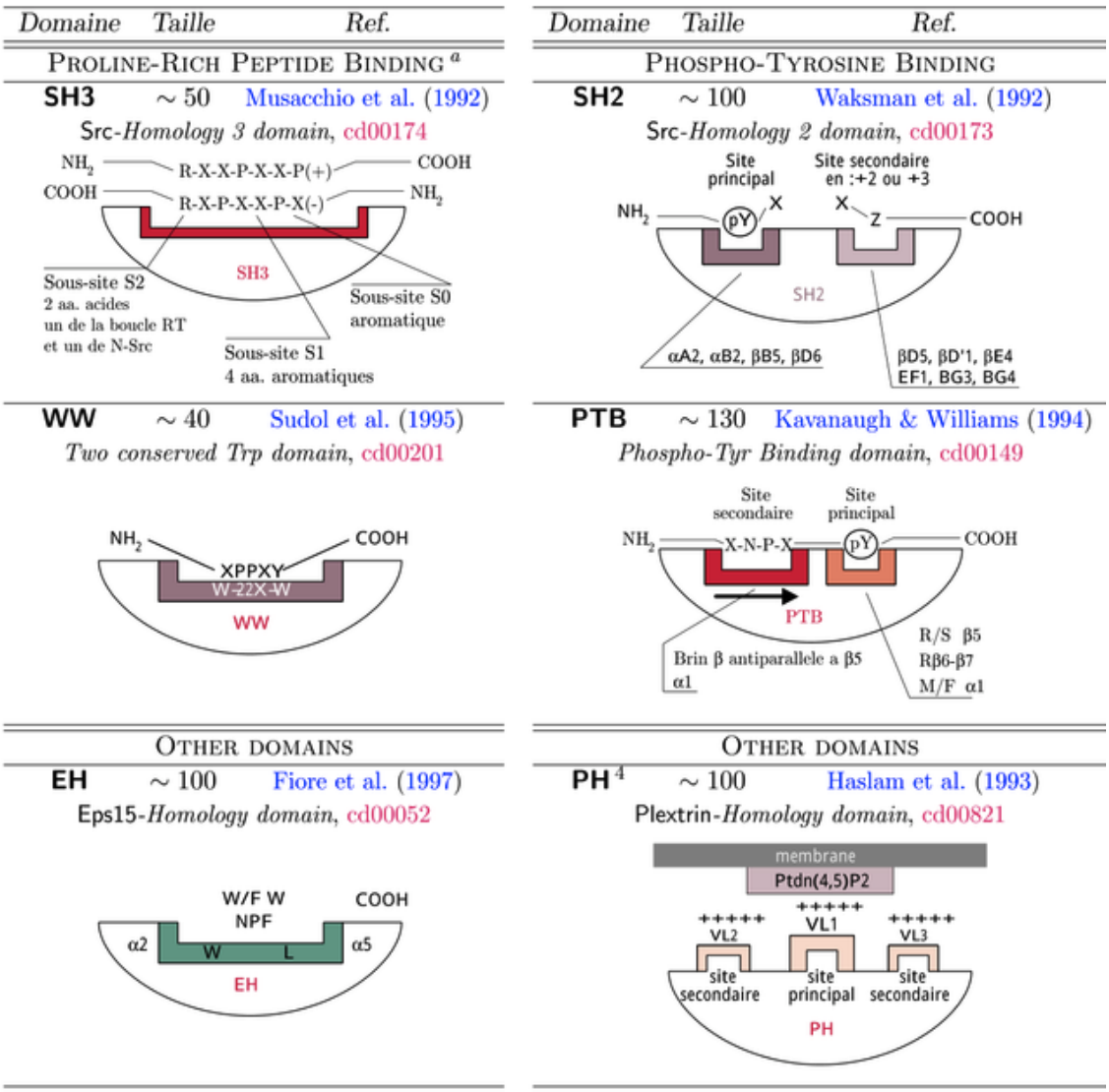


Figure 3.1 – Les principaux domaines d'interaction protéine-protéine (d'après Broutin & Ducruix [2000]). Pour chaque domaine, on précise son nom, sa taille en nombre de résidus et une référence bibliographique le décrivant. Sont décrits dans ce schéma : deux domaines (SH3 et WW) qui interagissent avec des peptides riches en Proline, deux domaines (SH2 et PTB) qui disposent d'une cavité pour accueillir une Tyrosine d'une autre protéine, et deux autres domaines utilisant d'autres mécanismes d'interaction.

modifiées par cette fixation. Il est également intéressant de remarquer que les domaines d'interaction présentés ici ont une taille relativement importante (50-100 acides aminés), alors que la longueur des peptides reconnus est souvent très faible (moins de 10 acides aminés).

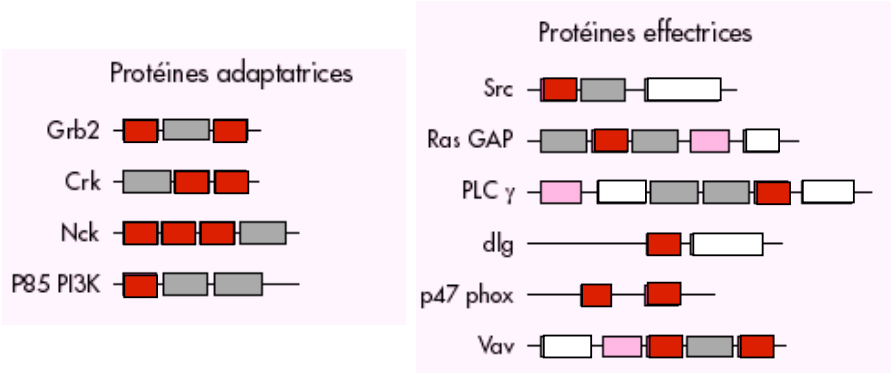


Figure 3.2 – Protéines de signalisation et domaines d'homologie. Les cadres gris, rouges, roses et blancs correspondent respectivement aux domaines SH2, SH3, PH (*Pleckstrin Homology*) et catalytiques des protéines.

3.1.2.1 Principes de la signalisation par interaction protéine-protéine

La compréhension du fonctionnement de cette transmission peut être complétée par plusieurs explications, notamment :

- Modification conformationnelle : lors de la fixation sur le domaine d'interaction, les deux partenaires vont très certainement subir des modifications conformationnelles, activant ou désactivant, par exemple leurs propriétés catalytiques respectives,
- Recrutement de protéines dans une localisation particulière de la cellule (par exemple à proximité des membranes nucléaires ou cellulaires, de chromosomes ou des centrosomes, d'organelles, ...). : la modification de l'emplacement d'une protéine peut en effet lui permettre d'interagir avec d'autres protéines ou structures cellulaires et donc de transmettre à nouveau des signaux.

Les interactions protéine-protéine par l'intermédiaire de domaines d'homologie donnent souvent naissance à des complexes multiprotéiques.

3.1.3 Rôles des domaines SH3

Les domaines SH3 ont un rôle dans le contrôle de l'architecture cellulaire et une fonction dans la localisation cellulaire des protéines. Par exemple, des études génétiques chez

la levure ont montré que des protéines à domaines SH3 comme ABP1, SLA1 et BEM1 sont nécessaires à l'organisation et à la polarisation du cytosquelette d'actine [Cohen *et al.* 1995 ; Pawson 1995]. De même, des mutations dans le gène *dlg* chez la drosophile, en particulier dans la région codant pour son domaine SH3, conduisent à une perte des jonctions serrées des cellules épithéliales [Cohen *et al.* 1995 ; Pawson 1995]. Par ailleurs, il a été montré que les domaines SH3 de Grb2 étaient nécessaires pour sa localisation au niveau des replis membranaires et celui de la PLCgamma pour sa fixation sur les microfilaments d'actine du cytosquelette [Cohen *et al.* 1995 ; Pawson 1995 ; Bar-Sagi *et al.* 1993].

Par coopération entre leurs domaines SH2 et SH3, les protéines adaptatrices ont un rôle d'intermédiaire dans le trafic protéique. Ainsi, Grb2, en agissant avec les récepteurs à activité tyrosine kinase par son domaine SH2 et avec le facteur d'échange Sos par ses domaines SH3, permet le recrutement rapide de Sos à la membrane. Celui-ci peut alors activer Ras et déclencher le processus de division ou de différenciation cellulaire [Fry *et al.* 1993 ; Pawson 1995 ; Lowenstein *et al.* 1992].

Les domaines SH3 régulent des activités enzymatiques. Certaines protéines, comme Grb2, la PI3K ou l'amphiphysine, interagissent par leur domaine SH3 avec la dynamine, une protéine essentielle pour l'endocytose des vésicules synaptiques en stimulant son activité GTPase [Scaife & Margolis 1997]. À l'inverse, certaines protéines tyrosine kinases, comme Abl, ont leur activité enzymatique régulée négativement par l'occupation de leur domaine SH3. Ceci pourrait expliquer l'activation oncogénique de la protéine lors de la délétion du domaine SH3 [Cicchetti *et al.* 1992].

Les domaines SH3 recrutent des substrats à leur enzyme. Ainsi, la protéine Src phosphoryle ses substrats AFAP-110 et Sam68 après les avoir recrutés par son domaine SH3. Dans les cellules phagocytaires, les domaines SH3 des protéines P40, P47 et P67 Phox jouent un rôle, à la fois dans l'assemblage et l'activation de la NADPH oxydase, ce qui leur attribue une fonction dans les réponses aux infections [Cohen *et al.* 1995 ; Pawson 1995].

Les domaines SH3 sont également impliqués dans certaines pathologies. La région riche en Prolines de la protéine Nef du HIV reconnaît le domaine SH3 des kinases Hck et Fyn avec une forte affinité [Kuryan & Cowburn 1997]. Dans certains cancers, la situation des domaines SH3, en aval de protéines tyrosine kinases oncogéniques, confère à des protéines de signalisation comme Grb2 un rôle de cible thérapeutique. Ces observations ont conduit à la recherche d'inhibiteurs des interactions passant par les domaines SH2 et SH3 [Smithgall 1995].

3.1.3.1 Spécificité de l'interaction

Le domaine SH3 d'une protéine peut interagir avec plusieurs cibles *in vivo* (Grb2 peut lier Sos, Vav, la dynamine...). De même, une protéine à région riche en Prolines peut se lier à différents domaines SH3 (la dynamine lie Grb2, la PI3K, l'amphiphysine). Ces observations, en relation avec les affinités faibles des séquences riches en Prolines pour les domaines SH3 (constante Kd de l'ordre de 10^{-5}M), posent le problème de la spécificité de ces interactions. Une première réponse est donnée par l'existence de deux classes de reconnaissance I et II, par la nature des chaînes latérales des acides aminés et par la localisation cellulaire des protéines. Néanmoins, les affinités et spécificités des protéines à domaines SH3 pour leurs cibles sont grandes [Kuryan & Cowburn 1997]. Ceci s'explique par l'existence de sites de reconnaissance additionnels, soit par l'intermédiaire de plusieurs domaines, soit sur un même domaine. Ainsi Vidal *et al.* ont montré que les acides aminés 36 à 45 du domaine SH3 N-terminal de Grb2 constituaient un second site de liaison avec certains de ses ligands, en complément de la conventionnelle plate-forme de reconnaissance des peptides riches en Prolines. En effet, Grb2 a deux domaines SH3 susceptibles d'interagir avec deux des régions riches en Prolines de Sos, tandis qu'une seconde région du domaine SH3 C-terminal de Vav est importante pour l'interaction avec la protéine nucléaire Ku [Romero *et al.* 1996b]. Cette région, très exposée, pourrait constituer un second site de reconnaissance pour certains domaines SH3, puisque la région

correspondante est également essentielle à l'interaction du domaine SH3 de GAP avec sa cible G3BP Yang *et al.* [1994].

3.2 Caractéristiques des domaines SH3

3.2.1 Homologie de séquence

Les domaines SH3 ont une faible homologie de séquence : parmi les 170 domaines présents dans les protéines humaines, après alignement de séquence, on ne trouve en moyenne que 30% de similarité.

La matrice présentée dans la figure 3.3 est issue de la base de données PROSITE. Elle représente pour chaque position de la séquence d'un domaine SH3, les acides aminés les plus fréquents. L'étude de cette matrice permet de conclure que seuls cinq acides aminés sont très conservés : une Tyrosine au début de la boucle R_t, deux Tryptophanes sur le brin S_3, une Glycine sur le brin S_4 et une Proline qui joue un rôle très important dans la boucle 3_{10} (qui est en fait un début d'hélice α). Sur cette figure, on trouve également la séquence consensus des domaines SH3, qui représente donc, pour chaque position de l'alignement de l'ensemble des séquences des SH3 connus, l'acide aminé le plus fréquemment rencontré. Étant donné la faible similarité de séquence entre ces domaines SH3, la pertinence des conclusions que l'on peut tirer de cette séquence consensus est relativement faible.

La figure 3.4 présente un alignement de différents domaines SH3 de protéines impliquées dans la signalisation cellulaire, en précisant les différents éléments de structure secondaire, ainsi qu'en indiquant les acides aminés les plus conservés.

3.2.2 Homologie de structure

La structure type d'un domaine SH3 est un tonneau β composé de deux feuillets β anti-parallèles. Les principales différences de structure entre les domaines SH3 que l'on

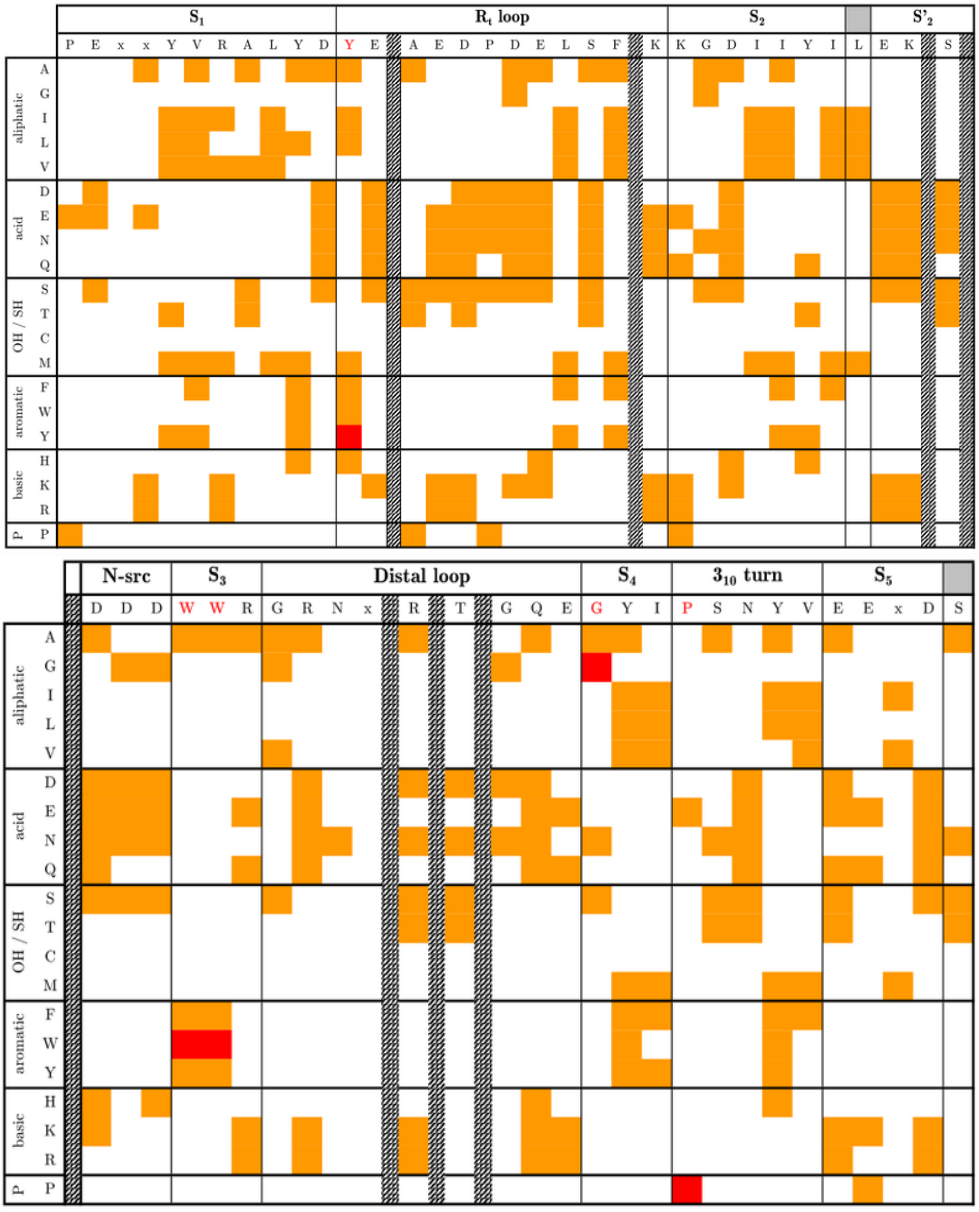


Figure 3.3 – Matrice simplifiée de définition du motif des domaines SH3. Le motif SH3 a la référence *PS*50002 dans la base de données PROSITE [Sigrist *et al.* 2002]. Cette matrice permet de calculer le score de n'importe quelle séquence protéique pour déterminer s'il s'agit d'un domaine SH3. Sur cette figure, les valeurs de score, obtenues statistiquement à partir d'un jeu initial de domaines reconnus comme étant de type SH3, ont été remplacées par un code de couleur pour une facilité de lecture. En haut de la figure est notée la séquence consensus. Pour chaque position au sein d'un domaine SH3, les acides aminés les plus fréquemment rencontrés sont notés en orange. On retrouve en rouge les acides aminés les plus conservés.

peut rencontrer, proviennent des différentes boucles reliant les brins β, dont les longueurs et positionnements sont très variables.

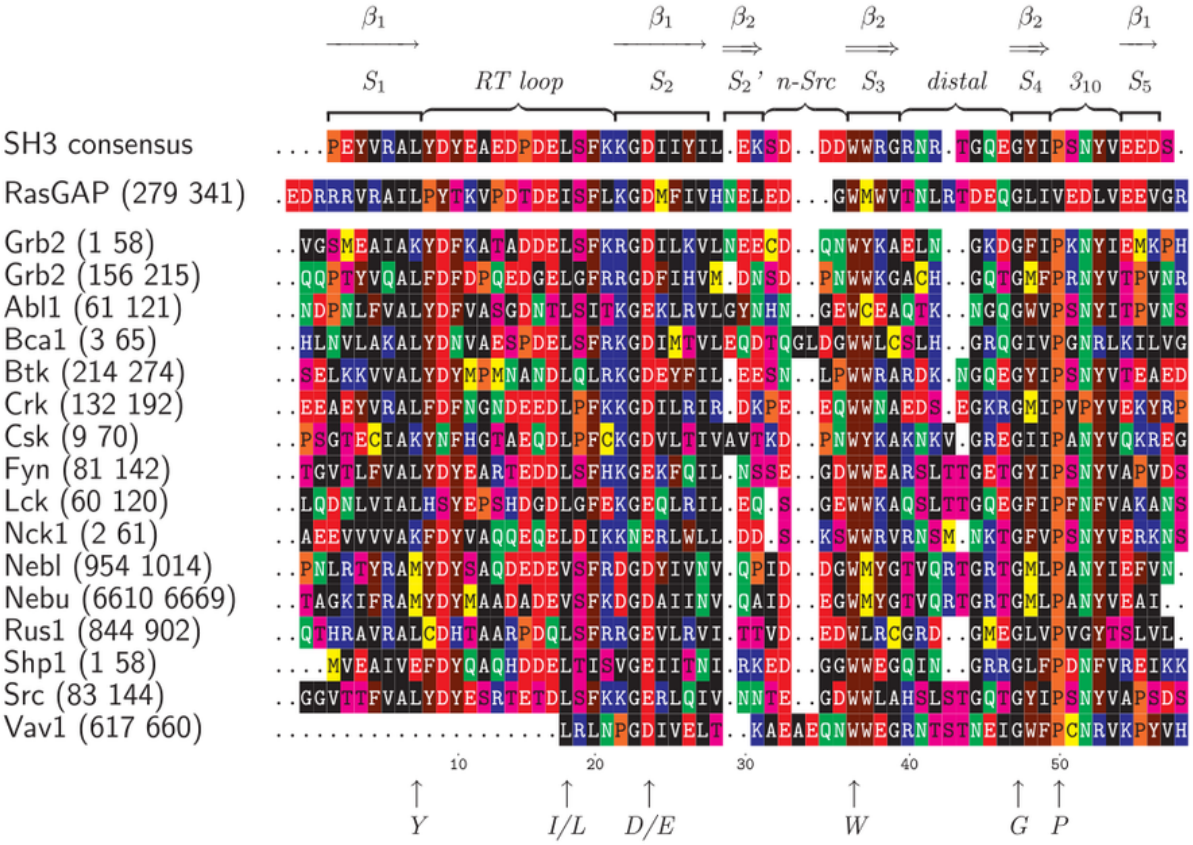


Figure 3.4 – Alignement de séquences des domaines SH3 de différentes protéines. La première ligne correspond à la séquence consensus de la définition des domaines SH3.

La topologie des domaines SH3 a été très bien caractérisée (figure 3.5). Le tonneau β est constitué de deux feuillets, notés ici β_1 et β_2. Le feuillet β_1 est lui même constitué de trois brins S_1, S_2 et S_5, et le second feuillet β_2 est composé du brin S_2' (qui prolonge sans rupture réelle le brin S_2 et des brins S_3 et S_4. Cette topologie particulière donne à ce domaine une très grande stabilité de par ce repliement presque semblable à un noeud.

La forte compacité des domaines SH3, leur importance dans la signalisation cellulaire, et leur faible taille en font des modèles de choix pour la modélisation moléculaire : de nombreux travaux de détermination de structures *ab initio* et de détermination de modes de repliements ont pris pour cible des domaines SH3. [Borreguero *et al.* 2004]

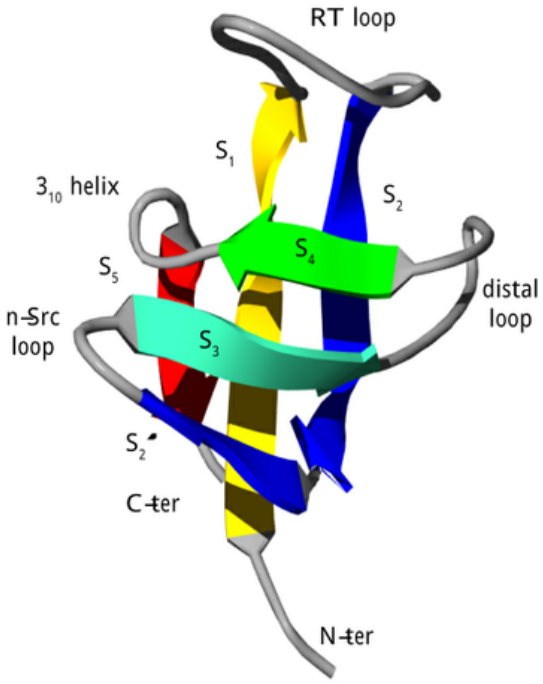


Figure 3.5 – Topologie du domaine SH3 de RasGAP (représentation schématisée des structures secondaires.

3.2.3 Ligands des domaines SH3

Les domaines SH3 interagissent conventionnellement avec des motifs de type hélice polyproline de type II (hélice PPII, [Zarrinpar *et al.* 2003]). Une hélice de ce type est une structure en hélice tournant à gauche, avec trois acides aminés par tour et une allure générale ressemblant à un prisme à base triangulaire. Ces hélices sont structurées par la présence en leur sein de plusieurs Prolines, dont la substitution de l'azote induit une contrainte structurale importante favorisant la conformation en hélice PPII [Creamer & Campbell 2002]. On peut voir une description schématisée de ce type d'hélice dans la figure 3.6.

Deux modes de reconnaissance de peptides riches en Prolines ont été identifiés :

– I : le peptide reconnu a son extrémité N-terminal au niveau de la boucle RT-loop du SH3. Sa séquence consensus est de type RxhPPhP,

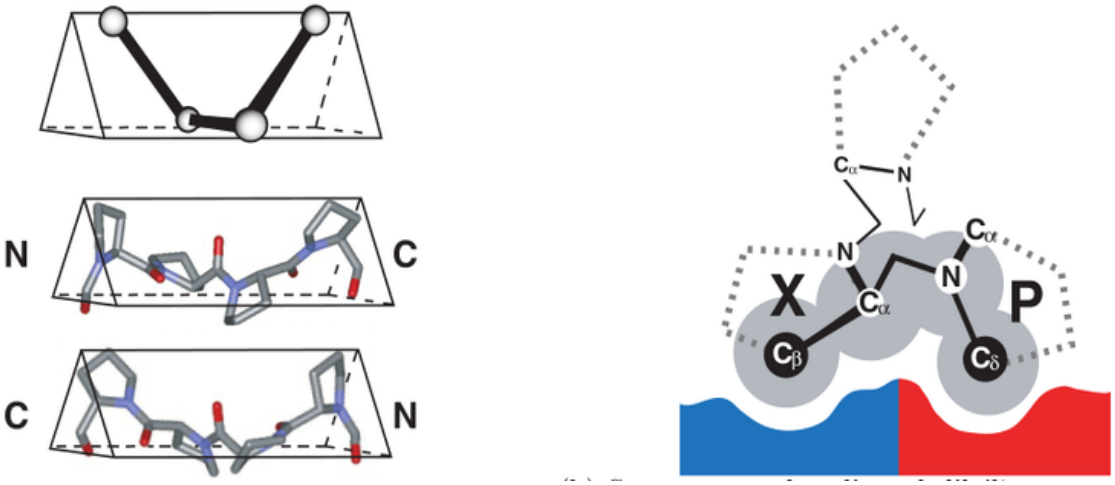


Figure 3.6 – Structure schématique des hélices polyprolines de type II. A gauche, une hélice PPII s'inscrit parfaitement dans un prisme à base triangulaire. A droite, on est dans l'axe de l'hélice et on peut noter parfaitement que les carbones β, ainsi que les substitutions des azotes de la chaîne principale jouent des rôles prépondérants dans la conformation de l'hélice et dans sa reconnaissance par un domaine SH3.

– II : le peptide reconnu a son extrémité C-terminale au niveau de la boucle RT-loop du SH3. Sa séquence consensus est de type PxhPxR.

Bien qu'un grand nombre de structures de domaines SH3 ait été résolu, en présence ou non de leurs ligands, les déterminants essentiels de la reconnaissance de ligands de classe I ou de classe II ne sont pas encore parfaitement identifiés. Certains domaines SH3 (par exemple, le domaine SH3 de Fyn) reconnaissent des ligands de chacune des deux classes. En revanche, il semble que d'autres domaines ne reconnaissent que des ligands de classe II. Fernandez-Ballester *et al.* [2004] ont proposé comme explication à cette spécificité de reconnaissance des différences de conformation du tryptophane très conservé.

Sur les domaines SH3 de Grb2, il a été très nettement identifié une plate-forme de reconnaissance pour ces peptides. Cette plate-forme comporte des poches de reconnaissance de deux Prolines et d'un acide aminé basique (Arginine souvent). La structure en hélice polyproline place les deux Prolines des positions i et $i+3$ d'un même côté de l'hélice et leur permet d'entrer dans deux poches adjacentes (notées sur la figure 3.7 S_0 et S_1) à la surface du domaine SH3.

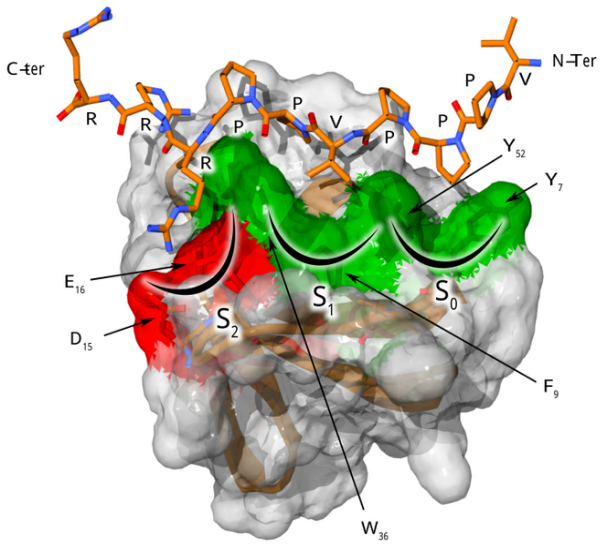


Figure 3.7 – Le domaine SH3 N-terminal de Grb2 complexé par le peptide VPPPVPPRRR (PDB 1AZE). [Vidal *et al.* 1998]

Il existe tout de même un certain nombre de domaines SH3 reconnaissant des ligands n'appartenant pas aux classes mentionnées ci-dessus. [Romero *et al.* 1996a ; Kang *et al.* 2000b ; Liu *et al.* 2003]

3.3 Intérêts de cette étude sur les domaines SH3

3.3.1 Conception d'inhibiteurs d'interactions protéine-protéine

La recherche d'inhibiteur d'interactions protéine-protéine est une approche relativement nouvelle dans le cadre de la conception de molécules thérapeutiques ciblées. Les revues de Pagliaro *et al.* [2004], de Berg [2003] ou de Ockey & Gadek [2002] présentent quelques interactions pour lesquelles de petites molécules ont été identifiées ainsi que de nouvelles techniques de criblage adaptées à cette problématique.

Néanmoins, la conception de petites molécules inhibitrices d'interactions protéine-protéine est rendue difficile par les importantes surfaces protéiques impliquées dans ces interactions.

Des résultats très encourageants, tant sur le plan de la découverte de petites molécules [Degterev *et al.* 2001] que de peptides et peptidomimétiques [Shangary & Johnson 2002 ; Walensky *et al.* 2004] ont été obtenus sur le modèle des interactions entre les domaines BH3 (Bcl-2-*Homology Domain 3* ou *Death domain*) et leurs partenaires.

3.3.1.1 Conception d'inhibiteurs de domaines SH3

L'importante surface d'interaction entre les domaines SH3 et leurs ligands rend très difficile la conception de petites molécules pouvant être utilisées comme molécules thérapeutiques. Cependant, une meilleur connaissance des caractéristiques des sites d'interaction impliqués dans ces reconnaissances devrait permettre dans certains cas d'identifier de petites molécules inhibitrices de domaines SH3.

Oneyama *et al.* [2002, 2003] ont identifié une petite molécule, UCS15A, capable d'inhiber l'interaction Sam68-Src, qui reconnait non pas le domaine SH3 de Src, mais la région riche en Proline de Sam68.

La recherche d'inhibiteurs non peptidiques de domaines SH3 a été aussi explorée. L'approche rationnelle, par modification de peptides issus de protéines connues pour leur interaction avec des domaines SH3, a pour l'instant été privilégiée. On peut notamment citer d'importants travaux de modification sélective des différents acides aminés de l'hélice polyproline du peptide issu de Sos reconnaissant les domaines SH3 de Grb2 [Nguyen *et al.* 1998, 2000]. Le principe de ces travaux consiste principalement à tenter d'améliorer à la fois l'affinité et la spécificité de l'inhibition de ces interactions par des peptoïdes (peptides dont certains acides aminés sont non naturels).

Feng *et al.* [1996] ont obtenu par chimie combinatoire plusieurs familles de modifications de ligands peptidiques du domaine SH3 de Src. L'inhibition des interactions des domaines SH2 et SH3 de la kinase Src est une voie thérapeutique d'intérêt dans le traitement de l'ostéoporose [Susva *et al.* 2000]. La recherche d'inhibiteur des deux domaines SH3

[Vidal *et al.* 2004] de la protéine Grb2 et également de son domaine SH2 [Liu *et al.* 1999] a permis d'identifier des molécules possédant des activités anti-tumorales intéressantes.

Les domaines SH3 constituent donc des cibles d'un type nouveau dans le cadre à la fois de la conception rationnelle et du criblage à haut débit de chimiothèques, dans l'objectif d'obtenir des inhibiteurs de la signalisation à visée thérapeutiques. De plus, leur forte compacité et leur petite taille devraient permettre le succès de telles démarches.

Chapitre 4

Mon travail de thèse : étude du modèle computationnel

4.1 Contexte de recherche

Grâce aux récents progrès technologiques et à l'arrivée des séquenceurs de nouvelle génération, la quantité de données génomiques croît exponentiellement. La figure 4.1 illustre la différence entre le nombre de structures disponibles dans la PDB et le nombre de séquences disponibles dans Swiss-Prot (base de séquences de protéines annotées). Nous pouvons voir que le nombre de séquences annotées croît exponentiellement. De nos jours (2013), la PDB compte plus de 80 000 structures de protéines, et la Swiss-Prot plus de 500 000 séquences.

Il existe donc, d'une part, un besoin important et toujours grandissant de connaître la fonction et la structure des protéines dans le processus d'annotation, et d'autre part, des méthodes expérimentales qui ne peuvent répondre à cette demande. C'est à cette étape qu'interviennent les méthodes bioinformatiques de prédiction de structure de protéines.

Afin de prédire la structure d'une protéine, deux voies sont possibles. Si une homologie avec une ou plusieurs protéines connues est détectée, ces protéines serviront de base à la fabrication d'un modèle 3D. Si aucune homologie n'est détectée, les méthodes dites *ab*

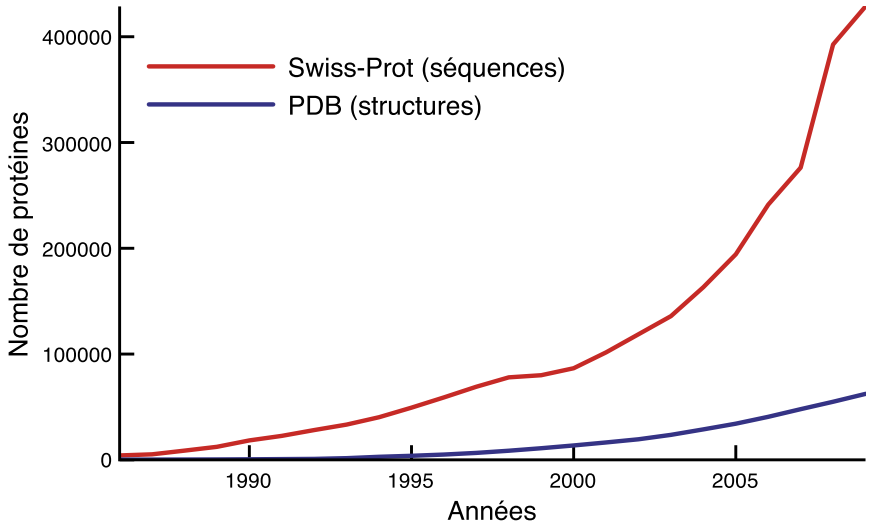


Figure 4.1 – Progression du nombre de séquences de protéines dans la banque de données Swiss-Prot et du nombre de structures de protéines dans la banque de données PDB entre 1986 et 2009.

initio sont utilisées pour construire un modèle 3D. La figure 4.2 illustre le diagramme méthodologique de la prédiction de la structure 3D d'une protéine.

Dans le mécanisme de détection d'homologie, la première étape consiste à aligner une séquence de protéine de structure inconnue avec des séquences de protéines connues. Si une forte similarité est détectée, il est très probable que celle-ci soit due à une relation d'homologie [Sander & Schneider 1991]. Néanmoins, en dessous de 30% d'identité de séquence, les mesures de similarité de séquences ne sont plus suffisantes pour détecter des homologies [Brenner *et al.* 1998 ; Rost 1999]. Il faut donc mettre en place d'autres méthodes afin de venir à bout de cette zone d'ombre.

Afin de détecter des protéines homologues dans la zone d'ombre, les méthodes dites de reconnaissance de repliements ont été développées dès 1990 [Hendlich *et al.* 1990 ; Sippl 1990], ainsi que les méthodes de repliement *ab initio*, et le CPD (*computational protein design*). Ces méthodes se basent sur la constatation que la structure des protéines est bien mieux conservée que leur séquence au cours de l'évolution [Illergard *et al.* 2009]. En effet, comme nous venons de le voir dans la partie « Introduction » de cette thèse, des protéines

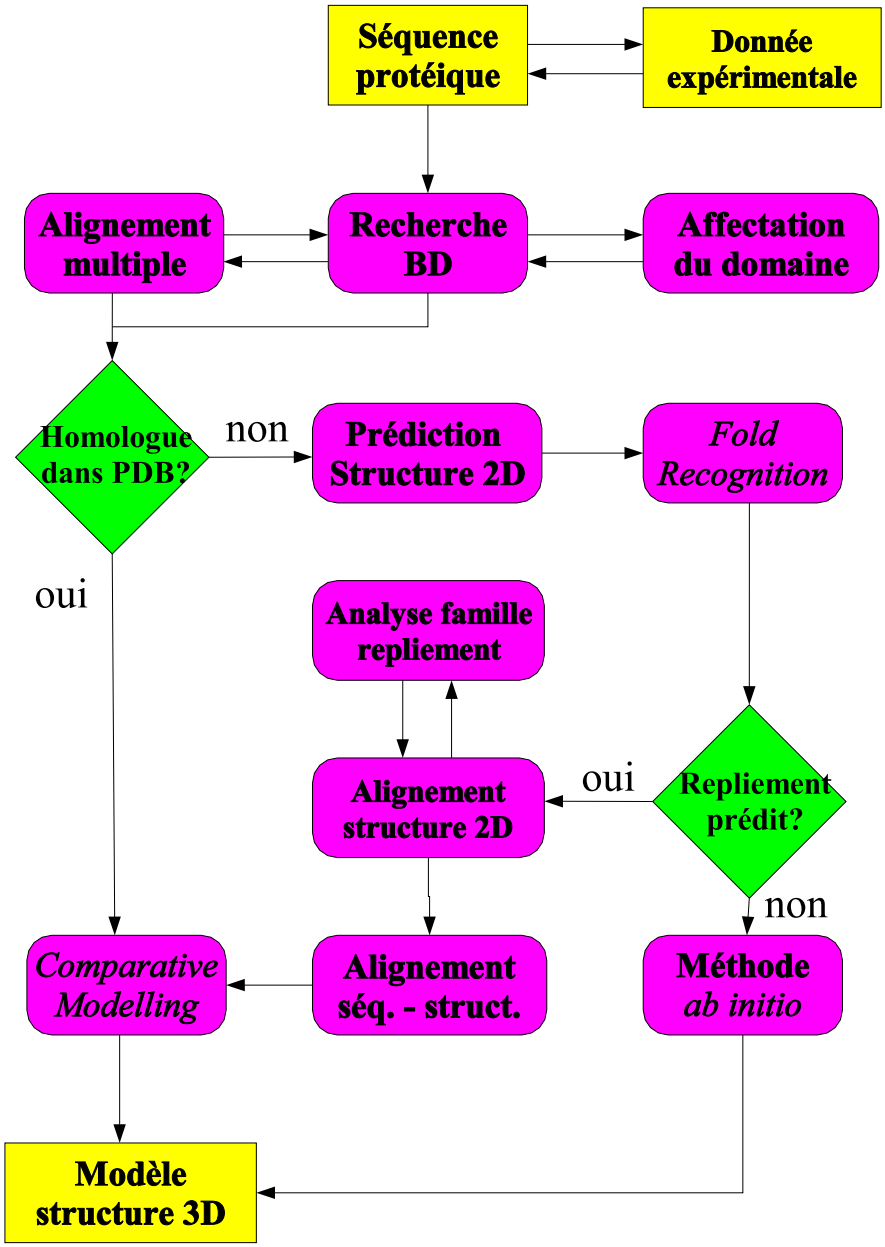


Figure 4.2 – Méthodologie pour la prédiction de la structure 3D d'une protéine.

ayant peu d'identité de séquences peuvent descendre d'un ancêtre commun et avoir une structure similaire. Il peut donc y avoir de nombreuses mutations dans une séquence de protéine sans que sa structure, voire sa fonction, ne soit changée.

Une question se pose alors : y a-t-il suffisamment de structures disponibles pour reconnaître n'importe quelle nouvelle séquence de protéine ? Zhang *et al.* [2006] on montré que les structures disponibles dans la PDB suffisent à reconnaître quasiment n'importe quelle protéine mono-domaine. De plus, de nombreuses études ont également montré que le nombre de repliements existants serait compris entre 1000 et 5000 [Chothia 1992 ; Wang 1996 ; Wolf *et al.* 2000 ; Coulson & Moult 2002 ; Liu *et al.* 2004]. De ces deux observations, nous pouvons en déduire que l'utilisation des structures résolues expérimentalement devrait nous permettre de classer la majorité des nouvelles protéines dans une famille existante. La réalité n'est bien sûr pas aussi simple.

Pour une structure donnée, on ne dispose souvent que d'une petite quantité de séquences natives y correspondant, et souvent assez peu identiques. Il est alors difficile de construire un profil de recherche d'homologues pour retrouver ces séquences dont on ne connaîtrait pas la structure. En effet, ces séquences étant trop éloignées d'un point de vue identité, la construction d'un profil moyennerait trop l'information en acides aminés. Alors comment disposer de bases de données de séquences plus conséquentes et exploitables pour chaque structure ?

Il a souvent été observé que des séquences très différentes peuvent avoir le même repliement. Ainsi, le design computationnel de protéine (CPD) considère le problème dans l'autre sens : si l'on connaît un repliement, est-il possible de retrouver la séquence ou l'ensemble de séquences qui lui correspondent ? Cependant, le nombre de séquences possibles pour une structure donnée est bien trop grand pour qu'il soit possible de générer toutes les séquences admissibles pour chaque repliement d'intérêt, et de les tester expérimentalement. Aussi, l'intérêt du CPD pourrait être ici de réduire la quantité des séquences possibles, et plus particulièrement dans cette thèse, de définir quelques séquences plus probables, à tester en priorité *in vitro*. La zone d'ombre pourrait être ainsi explorée en générant des bases de données de séquences théoriques correspondant à chacun des repliements connus.

Le protocole de génération de séquences théoriques correspondant à une structure donnée, développé par l'équipe de T. Simonson, nécessite plusieurs phases de validation et d'analyse à divers niveaux. C'est donc ici que ce situe le travail de ma thèse, plus précisément à l'interface de la bioinformatique théorique et de sa validation purement expérimentale :

- une analyse statistique et qualitative des séquences théoriques générées par CPD
- la mise en place d'une amélioration de la recherche d'homologue en utilisant ces séquences théoriques pour retrouver des séquences expérimentales non annotées
- la mise en place de critères et descripteurs pertinents pour choisir les meilleures séquences théoriques prédites
- la simulation par dynamique moléculaire des meilleures séquences théoriques choisies
- et l'étude expérimentale de ces séquences théoriques afin de confirmer la conservation de la structure 3D de départ.

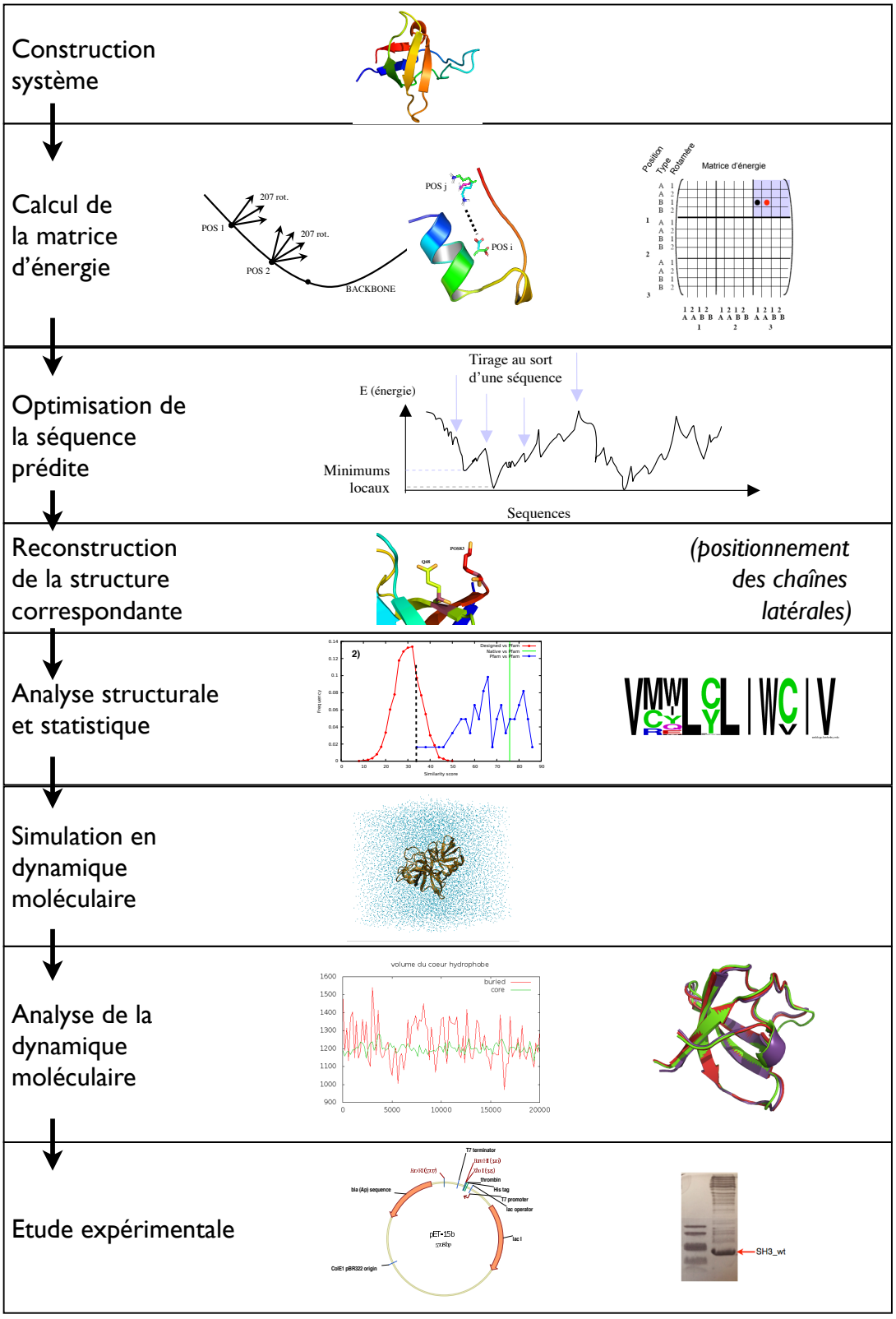


Figure 4.3 – Protocole général dans le cadre de cette thèse : de la génération de séquences théoriques à la sélection de candidats pour l'étude expérimentale.

4.2 Analyses statistiques sur les séquences théoriques et étude par homologie

J'ai tout d'abord mené à terme une première série d'analyses de covariance des sites d'alignements de séquences théoriques. Ainsi, grâce à notre processus de CPD, nous avons pu générer des séquences théoriques pour quelques structures des domaines SH2 et SH3. Nous disposons alors pour chaque structure étudiée, d'un alignement multiple de séquences théoriques sur lequel nous appliquons des calculs d'information mutuelle afin d'obtenir des informations sur les covariances entre acides aminés. À travers une analyse sophistiquée, nous tentons d'identifier les ensembles de positions les plus corrélées, que nous considérons comme réseau. A partir de quelques motifs fréquents d'acides aminés sur ces réseaux de positions, nous regroupons nos séquences théoriques en groupes. Nous pouvons ainsi faire des recherches d'homologues avec Blast pour chaque groupe. Nous moyennons donc moins l'information en réalisant plusieurs recherches sur de plus petits groupes de séquences respectant un motif, par rapport à l'information globale contenue dans l'alignement entier de séquences théoriques. Ainsi, nous pouvons identifier plus d'homologues avec ce protocole.

Néanmoins, la recherche d'homologue n'apportant pas suffisamment d'information sur la qualité des nos séquences théoriques, il convient de mettre en place un protocole de sélection des meilleures séquences. Puis de vérifier *in vitro* si ces séquences prédites peuvent être viables biologiquement et si leur structure sont bien identiques à celle de départ.

4.3 Analyses qualitatives des séquences théoriques

Afin de pouvoir étudier expérimentalement les séquences théoriques que nous avons prédîtes lors de la génération *in silico*, il fallait tout d'abord en sélectionner un nombre raisonnable et limiter le coût des expériences ensuite. Mais comment déterminer qu'une séquence théorique est meilleure qu'une autre ? Sur quels critères se baser pour les ca-

ractériser ? Comment choisir de l'ordre d'une dizaine de séquences théoriques de bonne qualité parmi 10 000 séquences de départ ? J'ai alors mis en place un nombre de descripteurs pour caractériser les séquences sur plusieurs critères, tels que la ressemblance avec la famille de protéines à laquelle appartiennent la protéine sauvage de départ, la ressemblance structurale avec la protéine sauvage initiale, ou divers comportements biologiques théoriques ...

Grâce à ces séries de filtres, nous pouvons nous focaliser sur plusieurs dizaines de séquences mutantes considérées de bonne qualité.

4.4 Analyse par dynamique moléculaire et choix de séquences théoriques à tester expérimentalement

Comme les générations de séquences théoriques par CPD se déroulent avec des structures assez fortement contraintes (squelette peptidique fixe), ainsi qu'une définition implicite du solvant, nous complétons les analyses sur nos prédictions structurales par des simulations de dynamique moléculaire avec un solvant explicite, afin de tester la stabilité de nos protéines mutantes.

Ainsi, le protocole de filtres précédent et l'analyse de stabilité par simulation de dynamique moléculaire nous ont permis de sélectionner de façon rationnelle une dizaine de séquences théoriques "candidates" en vue de les tester expérimentalement et de vérifier leur repliement *in vitro*.

4.5 Études expérimentales des séquences théoriques

Nous avons entrepris des tests d'expression et de purification sur la dizaine de séquences théoriques sélectionnées précédemment, afin d'étudier leur structure par des méthodes telles que le dichroïsme circulaire (CD) ou l'étude par RMN. Obtenir pour une

Code PDB	Méthod e	Résolution	Chaîne	Positions
1B07	X-ray	2.50	A	134-190
1CKA	X-ray	1.50	A	134-190
1CKB	X-ray	1.90	A	134-190
1JU5	NMR	-	B	217-228
1M30	NMR	-	A	134-190
1M3A	NMR	-	A	136-190
1M3B	NMR	-	A	136-190
1M3C	NMR	-	A	134-190
2GGR	NMR	-	A	230-304

Table 4.1 – Structures PDB connues de la protéine *Crk-mouse*. Nous pouvons y voir le code PDB, la méthode de résolution structurale (et sa qualité), et le détail de la séquence à considérer (chaîne et positions).

protéine mutante la même structure que celle de la protéine native initiale, permettrait à terme de valider notre modèle de prédiction de séquences par CPD.

4.6 Un domaine SH3 particulier pour l'étude du modèle computationnel

Constitués d'une soixantaine d'acides aminés, les domaines SH3 sont rencontrés dans un grand nombre de protéines de taille et de fonction variées : des protéines à activité enzymatique (tyrosine kinases cytoplasmiques, phospholipase Cgamma, ras-GAP pour *GTPase activating protein*), certaines protéines du cytosquelette (myosine, alpha spectrine), des protéines impliquées dans l'endocytose (amphiphysine), des facteurs d'échange (Vav), des protéines adaptatrices (Grb2, Crk, Nck, sous-unité 85 de la PI3-kinase).

Afin de réaliser toutes les étapes d'analyse sur ce modèle computationnel, nous avons choisi une structure particulière : la structure PDB 1CKA du domaine SH3 N-terminal de la protéine adaptatrice c-Crk (pour *Chicken Related Kinase*) de souris (notée *Crk-mouse*) dont le détail de la séquence est schématisé dans la figure 4.4. On liste les structures connues pour les domaines SH3 de cette protéine dans le tableau 4.1.

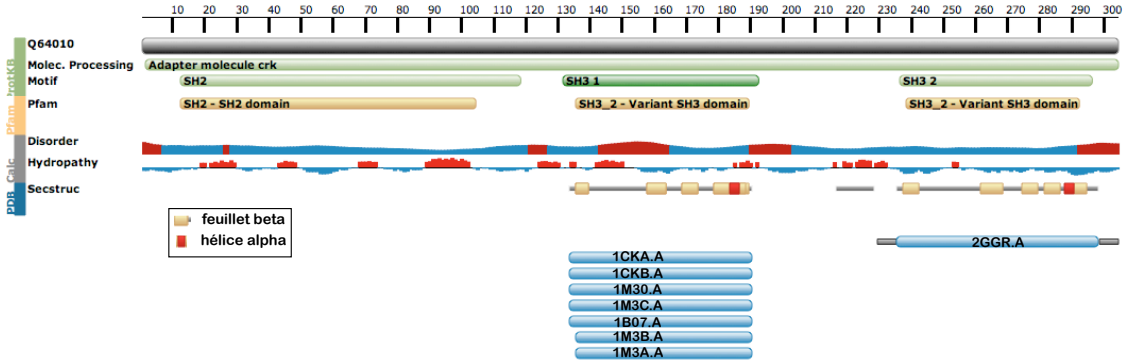


Figure 4.4 – Schéma de la protéine CRK adaptatrice.

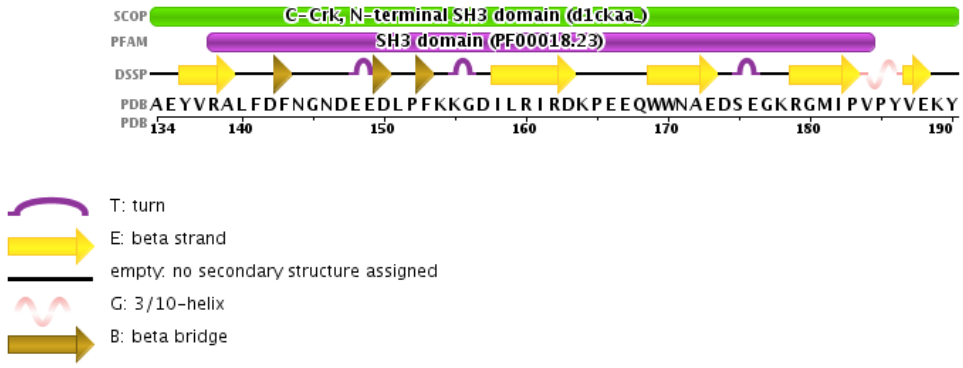


Figure 4.5 – Séquence annotée de la structure 1CKA du domaine SH3. (www.rcsb.org/pdb/)

Au cours de cette thèse, j'exposerai les résultats de mes recherches sur plusieurs protéines sauvage et mutantes pour la structure PDB 1CKA du domaine SH3 de la protéine c-Crk :

- la protéine sauvage du domaine SH3-1CKA, que nous noterons 1CKA-*wt*
- la protéine mutante du domaine SH3-1CKA dont la séquence a été générée par notre équipe avec un jeu de paramètres plus ancien, que nous noterons 1CKA-*old*
- des protéines mutantes du domaine SH3-1CKA dont les séquences ont été générées au cours de cette thèse en ne mutant pas les positions fonctionnelles sauvages (contact avec un ligand), puis filtrées avec un jeu de descripteurs, que nous noterons 1CKA-*lig*

– des protéines mutantes du domaine SH3-1CKA dont les séquences ont été générées au cours de cette thèse en ne mutant pas les positions du cœur hydrophobe sauvage, puis filtrées avec un jeu de descripteurs, que nous noterons 1CKA-*core*

Ainsi que les résultats sur d'autres protéines du domaine SH3-1CKA dont les séquences ont été générées avec d'autres paramètres afin de comparer leur efficacité grâce aux descripteurs mis en place.

Travail de thèse

Chapitre 5

Analyse statistique des séquences théoriques

Nous avons tenté de caractériser les séquences prédites par des méthodes bioinformatiques, notamment à travers leur utilisation pour la recherche d'homologues et la reconnaissance de plis. Ce travail est une collaboration avec mon co-directeur de thèse Jean-Marc Steyaert (LIX-Polytechnique).

La structure tridimensionnelle d'une protéine, et donc sa fonction biologique, dépend de sa séquence d'acides aminés. Les propriétés physico-chimiques des acides aminés induisent des interactions plus ou moins importantes entre eux, qui caractérisent et contraignent l'équilibre "mécanique" de la structure. La notion de covariance entre acides aminés découle naturellement de ces interactions structurales. La présence d'un résidu particulier à une position précise dans la séquence peut influencer le type d'acide aminé d'un autre résidu à une autre position, comme illustré simplement dans le schéma 5.1. On dira que ces résidus mutent ensemble, de façon corrélée. Ces covariances sont propres à chaque structure, c'est pour cela que nous avons voulu les étudier pour tenter de caractériser les familles de protéines.

La simulation informatique et la génération *in silico* de séquences théoriques vont nous permettre de trouver d'assez grands ensembles de séquences exploitables statistiquement correspondant à des structures particulières. Les modèles théoriques les plus

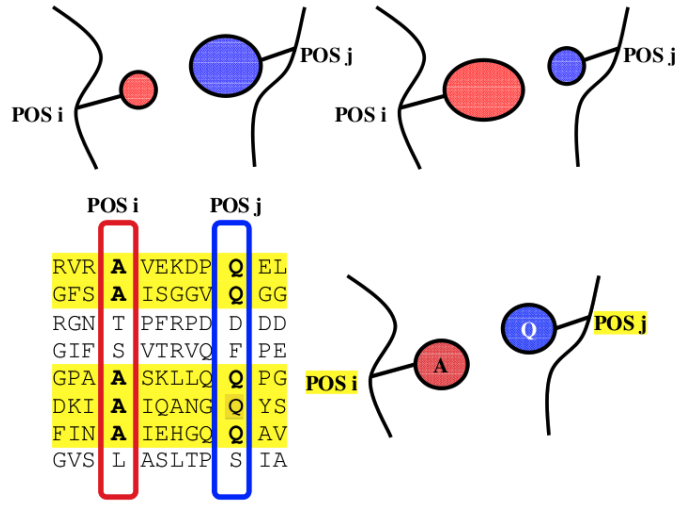


Figure 5.1 – Illustrations schématiques de covariances au sein d'un alignement de séquences et de la structure 3D. (En haut) Illustration simple sur la structure 3D de la possibilité d'avoir en position i un petit acide aminé et un gros en position j, ou la possibilité d'avoir en i un gros acide aminé et un petit en j. (En bas) Illustration simple de la notion de covariance. L'acide aminé A en position i influence fortement la présence de l'acide aminé Q en position j. On peut voir le motif ij en gras et en jaune les séquences "A Q" correspondant au motif. Le sous-alignement jaune est considéré comme un groupe.

importants aujourd'hui s'appuient sur une "mécanique moléculaire". Ils représentent la protéine comme un ensemble de particules sphériques, incompressibles (les atomes), reliés par des ressort et portant chacun une charge électrique. Une fois mise en place, la paramétrisation d'un tel modèle à l'aide de données expérimentales devient un outil puissant pour examiner le repliement et la stabilité des protéines. Schmidt am Busch *et al.* [2010] ont donc mis au point une méthode d'évolution dirigée pour prédire des séquences d'acides aminés susceptibles de se replier dans une architecture 3D donnée. En bref, pour une structure 3D donnée (un pli), on génère des séquences compatibles par un processus de mutations aléatoires et de sélection.

Nous disposons alors pour chaque structure étudiée, d'un alignement multiple de séquences théoriques sur lequel nous appliquons des calculs d'information mutuelle afin d'obtenir des informations sur les covariances entre acides aminés. À travers une analyse sophistiquée, nous tentons d'identifier les ensembles de positions les plus corrélées entre

elles, appelés réseaux. Nous pouvons alors repérer des motifs fréquents d'acides aminés sur ces positions.

Un de nos buts est de mettre au point une méthode qui prendrait en compte les informations sur ces covariances afin de mieux caractériser une structure particulière. En se basant sur les motifs obtenus, nous essayons de construire des sous-ensembles de séquences où l'information sur les covariances leur est propre. Finalement, nous tentons de valider notre analyse de covariance en procédant à des recherches d'homologues sur nos ensemble des séquences théoriques. Si nos motifs sont cohérents, alors nous devrions constater des améliorations sur les nombres d'homologues trouvés avec les groupes.

5.1 Calcul de corrélation

Grâce à la génération de séquence et pour chaque structure, nous disposons donc d'un alignement multiple (MSA) de 5000 séquences théoriques (de meilleures énergies PROTEUS) que nous allons pouvoir analyser statistiquement.

5.1.1 Entropie de site

Nous avons voulu connaître la variabilité en type d'acides aminés de chaque position ou site de nos MSA. Pour cela, nous avons eu recourt au calcul d'entropie. Ici, l'entropie d'un site i représente concrètement la diversité des acides aminés à la position i dans le MSA. Une entropie nulle correspond à un site conservé. Nous la calculons suivant la formule d'entropie de Shannon :

$$E_i = -\sum_x p_i^x \log(p_i^x) \tag{5.1}$$

où p_i^x est la probabilité de trouver l'acide aminé de type x à la position i dans le MSA.

5.1.2 Information mutuelle et matrice de covariance

Nous avons voulu ensuite connaître les dépendances entre positions du MSA pour déterminer les covariances au sein de la structure.

Considérons que les variables X et Y représentent la présence d'un type d'acides aminés dans deux colonnes du MSA. Alors l'information mutuelle du couple (X,Y) représente leur degré de dépendance au sens probabiliste. Informellement, on dit que deux variables sont indépendantes si la réalisation de l'une n'apporte aucune information sur la réalisation de l'autre. C'est la "distance" entre la distribution jointe $P(X,Y)$, et le produit des distributions $P(X)$ et $P(Y)$. Ici, l'information mutuelle $M_{i,j}$ mesure l'influence d'une position i dans le MSA sur une autre position j :

$$M_{i,j} = \sum_x \sum_y p_{i,j}^{x,y} \log \frac{p_{i,j}^{x,y}}{p_i^x p_j^y} \tag{5.2}$$

La double somme se calcule sur les types d'acides aminés x et y situés respectivement dans les colonnes i et j du MSA. Le terme $p_{i,j}^{x,y}$ représente la probabilité d'avoir en même temps le type x en position i et le type y en position j. p_i^x est la probabilité d'avoir le type x en position i. Même notation pour p_j^y. Les probabilités sont établies par simple comptage dans les colonnes i et j du MSA (probabilité nulle si un type d'acide aminé est absent d'une colonne). $M_{i,j}$ forme une matrice de covariance carré symétrique. Remarquons que le terme diagonal $M_{i,i}$, soit le cas particulier d'une covariance d'un site avec lui-même, est égal à l'entropie E_i du site i.

5.1.3 Classification des acides aminés

Dans nos calculs de covariances, nous avons choisi un alphabet plus restreint de 9 types d'acides aminés (au lieu de 20) pour "gommer" les mutations entre acides aminés similaires et simplifier la matrice. Ainsi, les résidus similaires n'influencent plus la variabilité des sites. Les acides aminés ont été classés d'après leurs propriétés physico-chimiques, en

Nom de la classe	Composition en acides aminés	Proriétés physico-chimiques
Groupe I (P)	PRO	non mutable dans le modèle
Groupe II (A)	ALA / SER / THR	chaîne latérale OH
Groupe III (G)	GLY	sans chaîne latérale, non mutable dans le modèle
Groupe IV (W)	TRP	aromatique, gros résidu
Groupe V (F)	PHE / TYR	aromatiques
Groupe VI (H)	HIS	basique, aromatique
Groupe VII (R)	ARG / LYS	basiques
Groupe VIII (D)	ASP / ASN / GLU / GLN	acides et polaires
Groupe IX (L)	LEU / VAL / ILE / MET / CYS	hydrophobes

Table 5.1 – Alphabet réduit : répartition des 20 acides aminés en 9 groupes. Dans la première colonne, entre parenthèses, on trouve le code à une lettre représentation le groupe.

s'appuyant sur une analyse par "clusterisation" de la matrice BLOSUM62. Les classes d'acides aminés sont indiquées dans le tableau 5.1.

5.2 Analyse spectrale de la matrice de covariance

On applique à la matrice de covariance une analyse spectrale, classique en analyse numérique. La méthode consiste à diagonaliser la matrice de covariance pour obtenir les vecteurs propres. Chaque vecteur, appelé "mode", représente un "mouvement" d'ensemble d'acides aminés dans un espace inhabituel : "l'espace des mutations". Un mouvement correspond à un ensemble de résidus qui mutent ensemble et de façon corrélée. On s'intéresse surtout aux modes de plus grande amplitude, c'est-à-dire avec les plus grandes valeurs propres.

Ainsi, au lieu d'analyser la matrice complète, on analyse une matrice approchée dans laquelle on "superpose" les n premiers modes (dans l'ordre décroissant des valeurs propres). On reconstruit la matrice de covariance à partir de ces n modes grâce à l'équation :

$$M_{i,j}^{(n)} = \sum_{k=1}^{n} \lambda^k W_i^k W_j^k \tag{5.3}$$

où λ^k est la $k$ème valeur propre, W_i^k le $i$ème composant du $k$ème vecteur propre.

Pour choisir n, on va s'intéresser à la somme des valeurs propres λ^1, ..., λ^n et sa valeur relative à la somme complète de λ^1, ..., λ^N. Leur ratio mesure la proportion de la covariance totale qui est prise en compte.

Enfin, on ne garde qu'une matrice partielle $M^{(n)}$ en ne conservant que les composants supérieurs à un pourcentage (seuil) de la plus grande composante de la matrice $M^{(n)}$. Ce "seuillage" permet d'éliminer les petits éléments de la matrice.

5.3 Recherche de motifs

À partir de cette matrice de covariance "débruitée", nous cherchons à faire ressortir plusieurs positions dans le MSA qui covarient ensemble.

5.3.1 Sélection des positions du réseau

On analyse la matrice "débruitée", correspondant aux n modes superposés, avec une méthode heuristique intégrant quelques notions biologiques et bioinformatiques.

Après avoir visualisé sous Pymol toutes les positions qui contribuent principalement à la matrice (environ 10 à 15), on élimine celles situées dans des boucles de la structure 3D de la protéine. En effet, la méthode de génération de séquences traite les boucles comme des objets fixes, ce qui oblige à considérer les mutations observées avec prudence. On élimine aussi les positions "isolées".

Les positions sélectionnées sont considérées comme réseau. Au sein du MSA, ces positions vont adopter différents types d'acides aminés de manière corrélée. Certaines combinaisons d'acides aminés sont plus fréquentes que d'autres : nous nous intéresserons à ces motifs particulièrement.

5.3.2 Regroupement des séquences en fonction du motif

Les motifs obtenus ci-dessus vont permettre de regrouper les séquences théoriques en plusieurs sous-ensembles. On réduit ici de nouveau l'espace des types d'acides aminés pour effacer un peu plus les distinctions entre acides aminés similaires. Nos séquences théoriques sont donc préalablement transcrites dans un alphabet de 9 acides aminés avant de les trier avec un ordre lexicographique sur les positions du motif. On ne garde que les dix plus fréquentes séquences induites par le motif, ce qui donne 10 groupes de séquences.

5.4 Recherche d'homologues à l'aide de séquences théoriques

Nous allons maintenant soumettre nos sous-ensembles (ou groupes) de séquences théoriques construits précédemment, à des requêtes Blast. Lors de ces recherches d'homologues, des profils sont construits à partir des séquences de chaque groupe. En parallèle, nous effectuons une requête Blast construisant un profil avec toutes les séquences théoriques. Si l'utilisation des groupes améliorent le nombre d'homologues trouvés par rapport à la requête avec l'ensemble des séquences, cela nous permettra de valider notre analyse de covariance, ainsi que de mieux appréhender la qualité des séquences prédites.

5.4.1 Recherche d'homologues sans groupe

La recherche d'homologues à partir d'un ensemble S de séquences théoriques générées pour une structure donnée, se déroule en plusieurs étapes :

– Tirage aléatoire de 100 séquences parmi les 5 000 séquences théoriques de meilleure énergie de l'ensemble S.

- Requête Blast : construction d'un profil de recherche à partir de ces 100 séquences et recherche d'homologues dans une base de données de 772 domaines SH3 connus avec un seuil e-value de 0,001 et la matrice Blosum62
- 100 itérations sur ces 2 premières étapes, on considère ensuite l'ensemble de tous les homologues trouvés
- Élimination de la redondance et gestion de la transitivité dans l'ensemble des homologues : requêtes Blast à partir de cet ensemble d'homologues sur lui-même ; on regroupe les homologues qui seraient identiques à plus de 90%.

5.4.2 Recherche d'homologues avec groupes

L'analyse des covariances nous permet de travailler sur plusieurs ensembles de séquences plus petits (souvent moins de 100 séquences), afin de moins moyenner l'information au sein du MSA et de trouver ainsi plus d'homologues. Nous utilisons donc 10 sous-ensembles de séquences théoriques ayant un des 10 plus fréquents motifs sur les positions choisies, et l'ensemble des séquences restantes. La recherche d'homologues se déroule en plusieurs étapes :

- Pour chaque sous-ensemble de séquences (10 groupes + le reste)
- Requête Blast : construction d'un profil de recherche à partir des séquences d'un sous-ensemble et recherche d'homologues dans la base de données de 772 domaines SH3 connus avec un seuil e-value de 0,001 et la matrice Blosum62
- On considère ensuite l'ensemble de tous les homologues trouvés pour tous les sous-ensembles
- Élimination de la redondance et gestion de la transitivité dans l'ensemble des homologues : requêtes Blast à partir de cet ensemble d'homologues sur lui-même ; on regroupe les homologues qui seraient identiques à plus de 90%.

5.5 Résultats

5.5.1 Analyse des covariances et recherches d'homologues pour le domaine SH3-1CKA

5.5.1.1 Données

Nous avons voulu réaliser cette étude des covariances sur les séquences théoriques générées à partir du squelette peptidique de la protéine sauvage 1CKA, dans le cas où les positions fonctionnelles (en contact avec le ligand) ont été fixées dans leur type natif. Cette génération de séquences est détaillée dans le chapitre suivant sous la dénomination de 1CKA-*LIG*. Nous n'avons pris en compte que les 5 000 premières séquences de meilleure énergie PROTEUS.

Ensuite, on a réalisé deux versions du calcul de la matrice de covariance sur le même ensemble de séquences théoriques 1CKA-*LIG* : l'une avec un alphabet classique de 20 acides aminés (Tableau 1.2), l'autre avec un alphabet réduit de 9 groupes d'acides aminés. Cette classification des acides aminés en 9 groupes est détaillée dans le tableau 5.1.

5.5.1.2 Analyse des covariances sur les séquences avec alphabet de 20 acides aminés

Nous avons tout d'abord analysé la matrice de covariance calculée avec un alphabet complet de 20 acides aminés. Nous pouvons voir dans la matrice de covariance complète (figure 5.2) les structures secondaires représentées par des flèches bleues (feuillets beta) et des rectangles rouges (hélices alpha) sur les axes, facilitant ainsi la lecture des résultats. Néanmoins, nous pouvons vite remarquer des corrélations fortes entre certaines positions qui ne font pas partie de structures secondaires. Cela n'est que très peu plausible. En effet, on imagine bien que dans des boucles très variables en terme de composition en acides aminés, les corrélations devraient être rares. Mais nous avons introduit ce biais lors de la

Positions Num. à partir de 1	Num. Pymol à partir de 134	Valeur propre
2, 3, 31, 35, 42, 43, 45, 51	135, 136, 164, 168, 175, 176, 178, 184	3.3044
2, 3, 31, 35, 42, 43, 45, 51	135, 136, 164, 168, 175, 176, 178, 184	2.80393
2, 3, 25, 27, 29, 31, 40, 55	135, 136, 158, 160, 162, 164, 173, 188	2.53041
2, 3, 7, 9, 11, 13, 31, 55	135, 136, 140, 142, 144, 146, 164, 188	2.30337
5, 7, 9, 13, 14, 16, 35, 55	138, 140, 142, 146, 147, 149, 168, 188	2.02148
2, 9, 11, 14, 31, 35, 51, 55	135, 142, 144, 147, 164, 168, 184, 188	1.76389
6, 13, 14, 20, 25, 26, 35, 51	139, 146, 147, 153, 158, 159, 168, 184	1.64944
3, 9, 14, 25, 35, 38, 46, 51	136, 142, 147, 158, 168, 171, 179, 184	1.63356
3, 13, 14, 25, 35, 38, 46, 51	136, 146, 147, 158, 168, 171, 179, 184	1.5852
1, 25, 27, 29, 30, 38, 40, 46	134, 158, 160, 162, 163, 171, 173, 179	1.33631

Table 5.2 – Positions des 10 premiers modes propres tirées de l'analyse spectrale de la matrice de covariance avec alphabet de 20 acides aminés.

génération de séquences, en considérant le *backbone* fixe, et donc les boucles, induisant des contacts biaisés entre les positions des boucles et d'autres régions de la protéine. Nous nous sommes donc focalisées sur les positions dans ou proches des structures secondaires afin d'éviter le problème.

Dans la liste des positions des 10 premiers modes propres (figure 5.2), nous pouvons déjà observer des positions qui apparaissent fréquemment comme des "réseaux" de positions covariantes. En superposant plusieurs modes, nous tentons alors de rassembler des informations complémentaires sur ces "réseaux" pour qu'il en ressorte les principales covariances. Ainsi, en superposant les 4 premiers modes propres, nous conservons 30% de l'information totale. Néanmoins, pour ne considérer que les plus fortes corrélations, il convenait de ne prendre que les grandes valeurs propres, nous gardons ainsi une matrice avec 18% de l'information totale.

De ces 4 modes propres superposés s'est dégagé un ensemble de positions cohérentes : 135-136-158-160-164-188, qui se trouvent proches spatialement dans la structure 3D de la protéine, comme nous pouvons le voir dans la figure 5.6.

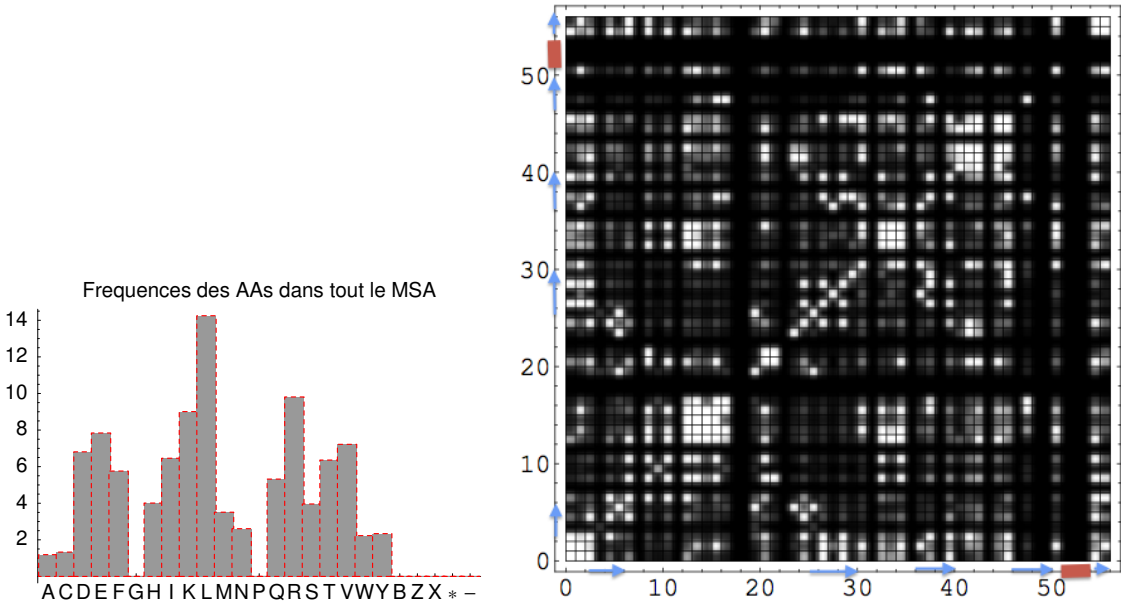


Figure 5.2 – Fréquence des acides aminés dans l'alignement de séquences théoriques. Matrice de covariance calculée avec l'alphabet de 20 acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

Modes pris en compte	Trace	Pourcentage
Matrice de covariance entière	35.6213	100.0 %
Modes 1 a 4	10.9421	30.8%
Modes 1 a 4 seuillés à 35%	6.68122	18.8%

Table 5.3 – Traces de la matrice de covariance entière et des modes propres 1 à 4 superposés tirées de l'analyse spectrale de la matrice de covariance avec alphabet de 20 acides aminés.

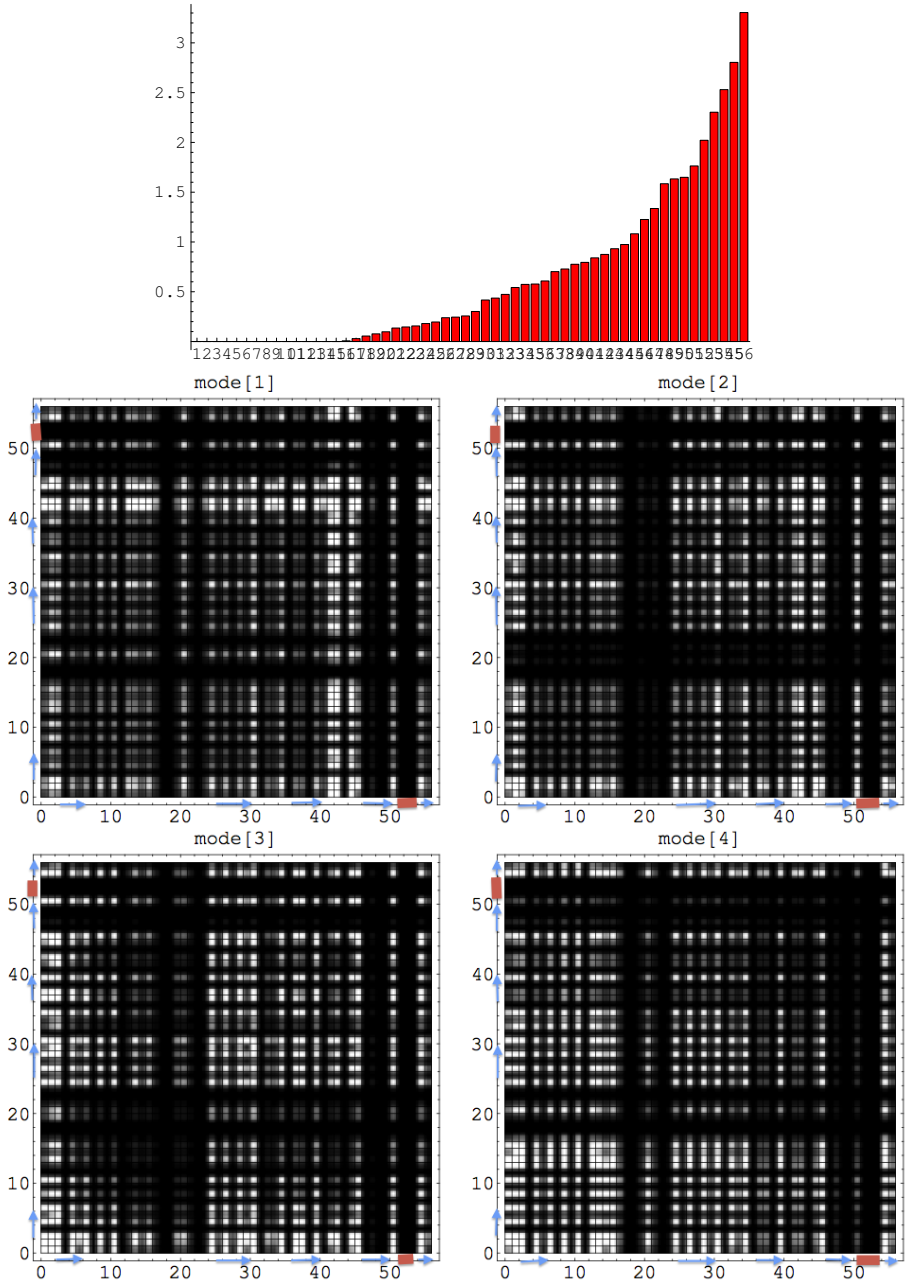


Figure 5.3 – Distribution des valeurs propres et matrices représentant les modes propres de 1 à 4 tirés de l'analyse spectrale de la matrice de covariance totale avec alphabet de 20 acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

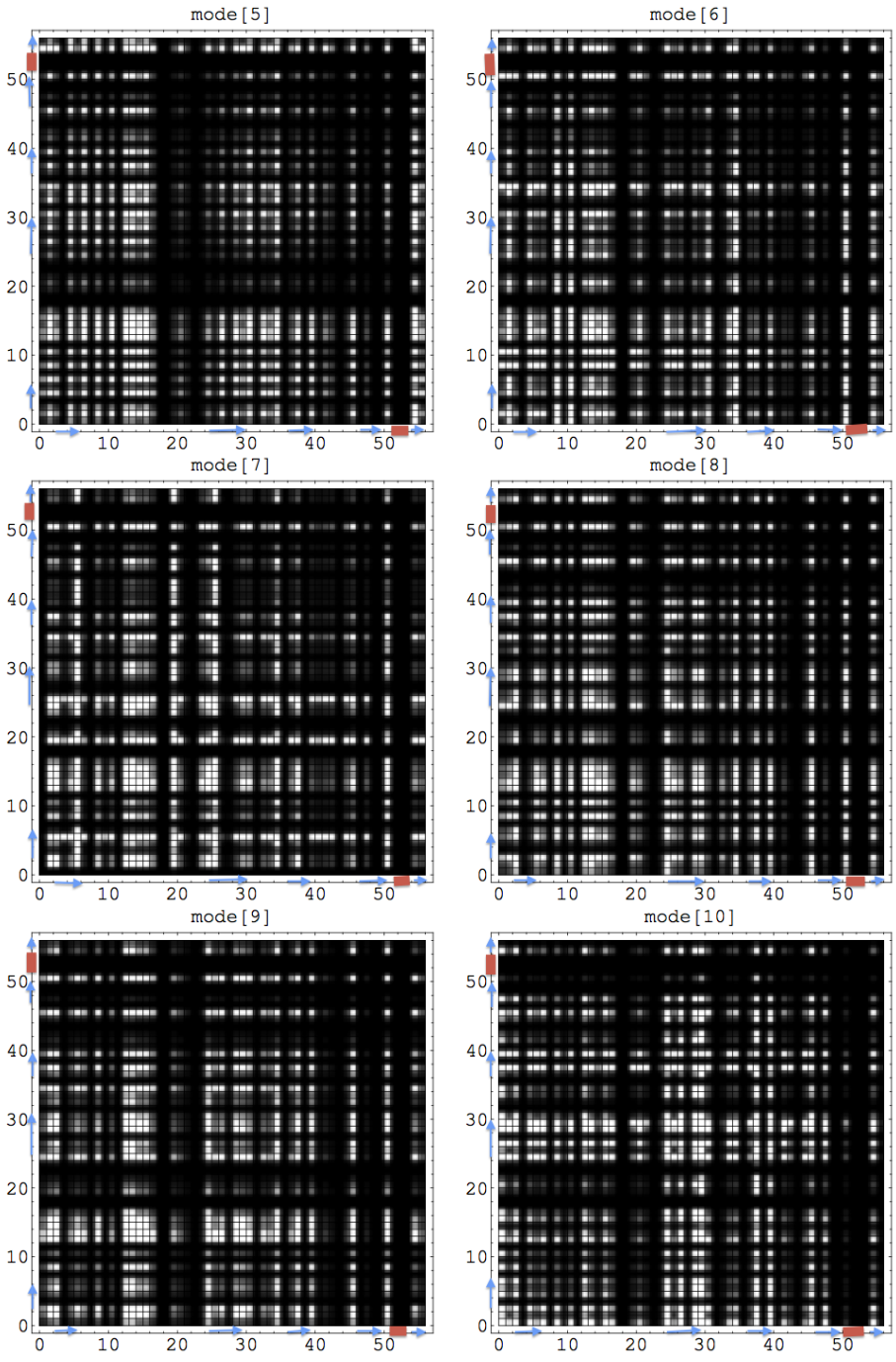


Figure 5.4 – Matrices représentant les modes propres de 5 à 10 tirés de l'analyse spectrale de la matrice de covariance total avec alphabet de 20 acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

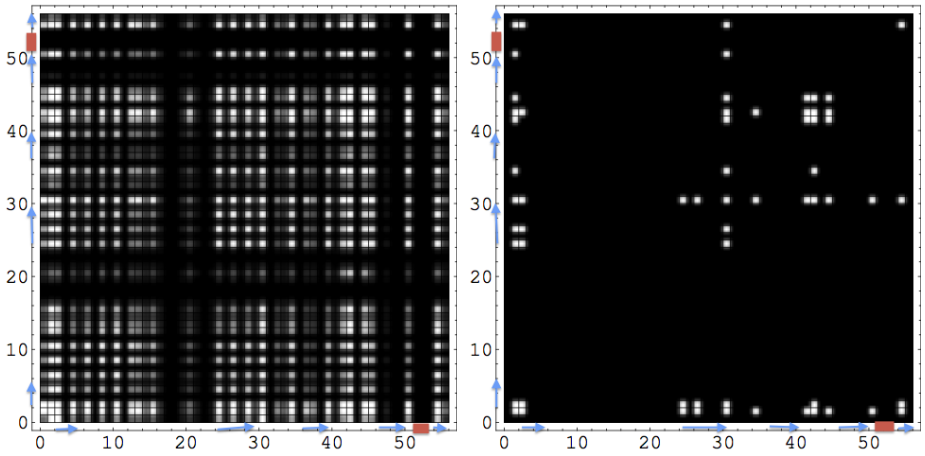


Figure 5.5 – Matrices représentant les modes propres 1 à 4 superposés et seuillé à 35 % tirés de l'analyse spectrale de la matrice de covariance totale avec alphabet de 20 acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

Figure 5.6 – Les positions 135-136-158-160-164-188 tirées de l'analyse spectrale de la matrice de covariance avec alphabet réduit d'acides aminés sont représentées par des sphères disposées sur le squelette peptidique de la protéine

Positions Num. à partir de 1	Num. Pymol à partir de 134	Valeur propre
2, 21, 31, 35, 42, 43, 45, 46	135, 154, 164, 168, 175, 176, 178, 179	2.49306
2, 3, 27, 29, 31, 35, 51, 55	135, 136, 160, 162, 164, 168, 184, 188	1.963
2, 3, 27, 29, 31, 40, 46, 55	135, 136, 160, 162, 164, 173, 179, 188	1.81872
3, 5, 7, 27, 29, 40, 46, 55	136, 138, 140, 160, 162, 173, 179, 188	1.75559
2, 14, 16, 31, 33, 35, 51, 55	135, 147, 149, 164, 166, 168, 184, 188	1.48011
9, 11, 13, 14, 16, 33, 35, 51	142, 144, 146, 147, 149, 166, 168, 184	1.2764
9, 11, 14, 16, 30, 35, 38, 46	142, 144, 147, 149, 163, 168, 171, 179	1.08296
3, 9, 14, 16, 30, 35, 38, 46	136, 142, 147, 149, 163, 168, 171, 179	1.06959
2, 9, 16, 31, 33, 35, 37, 51	135, 142, 149, 164, 166, 168, 170, 184	0.949633
3, 16, 29, 30, 33, 38, 40, 46	136, 149, 162, 163, 166, 171, 173, 179	0.909761

Table 5.4 – Positions correspondant aux 10 premiers modes propres tirées de l'analyse spectrale de la matrice de covariance avec alphabet réduit d'acides aminés.

5.5.1.3 Analyses des covariances sur les séquences avec alphabet réduit d'acides aminés

Ensuite, nous avons analysé la matrice de covariance calculée avec l'alphabet réduit de 9 acides aminés.

Dans le détail des 10 premiers modes propres (figure 5.4), beaucoup des positions sorties dans le cas de l'alphabet complet d'acides aminés apparaissent ici également.

Afin de comparer les méthodes, nous avons superposé les 4 premiers modes propres. Nous conservons ainsi 31% de l'information totale. Et même en considérant des corrélations plus fortes qu'avec l'alphabet 20 acides aminés (seuil moins tolérant, 20% au lieu de 35%), nous gardons ainsi une matrice avec 26,5% de l'information totale. Cela reste cohérent puisque la variabilité en acides aminés dans chaque position du MSA est gommée par l'utilisation de la classification des acides aminés. Ainsi, certaines covariances sont renforcées.

De ces 4 modes propres superposés s'est dégagé un ensemble de positions proches de celles trouvés avec l'alphabet complet : 135-136-160-162-164-188, elles ne différent que d'une position (162 au lieu de 158), comme nous pouvons le voir dans la figure 5.11.

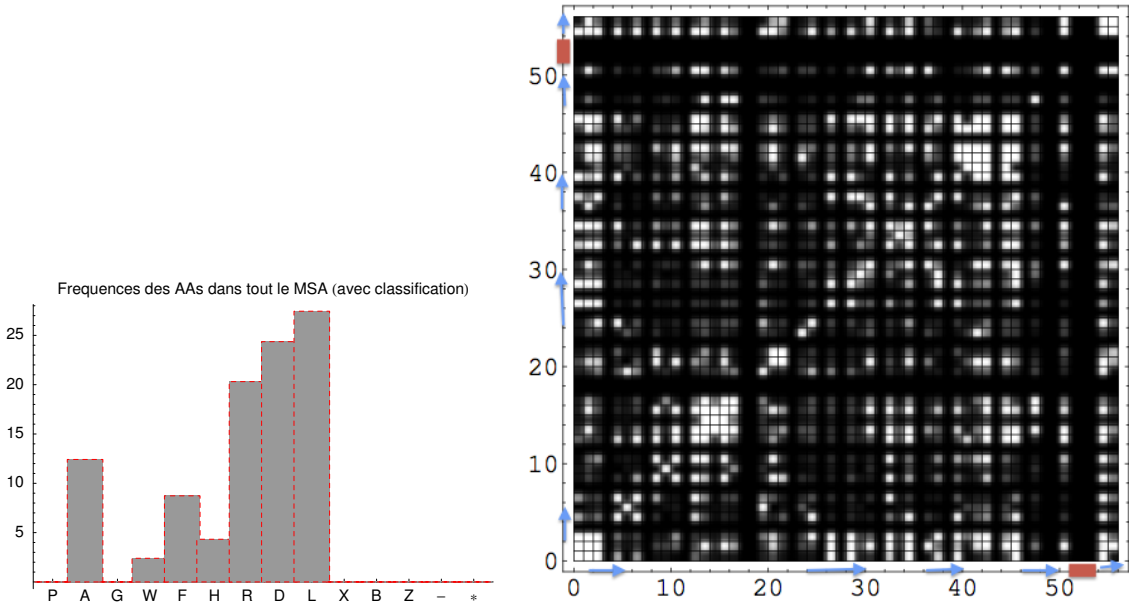


Figure 5.7 – Fréquences des groupes d'acides aminés dans l'alignement de séquences théoriques. Matrice de covariance calculée avec l'alphabet réduit d'acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

Modes pris en compte	Trace	Pourcentage
Matrice de covariance entière	25.6613	100.0%
Modes 1 a 4	8.03037	31.3%
Modes 1 a 4 seuillés à 20%	6.78052	26.5%

Table 5.5 – Traces de la matrice de covariance entière et des modes propres 1 à 4 superposés tirées de l'analyse spectrale de la matrice de covariance avec alphabet réduit d'acides aminés.

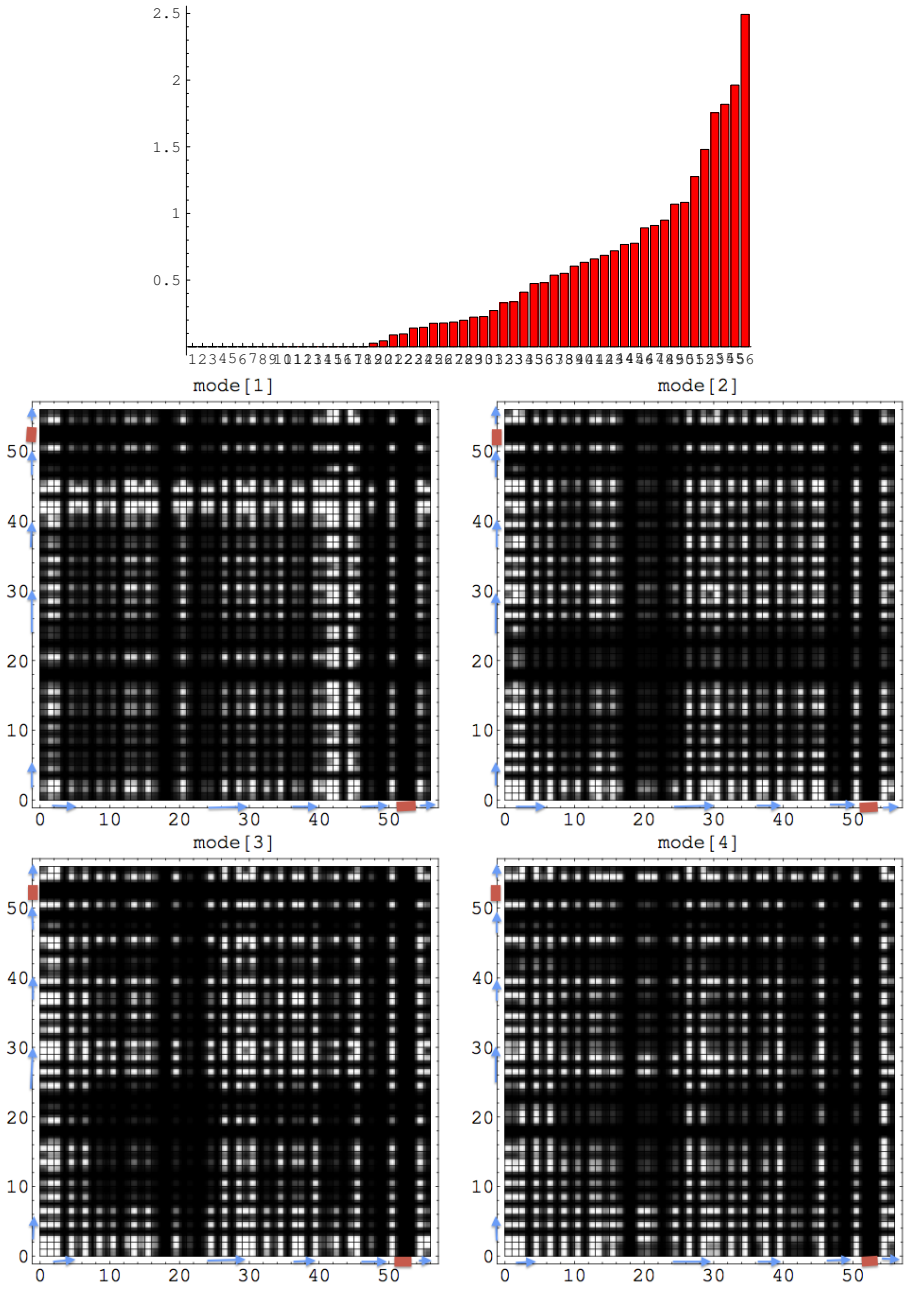


Figure 5.8 – Distribution des valeurs propres et matrice représentant les modes propres de 1 à 4 tirés de l'analyse spectrale de la matrice de covariance totale avec alphabet réduit d'acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

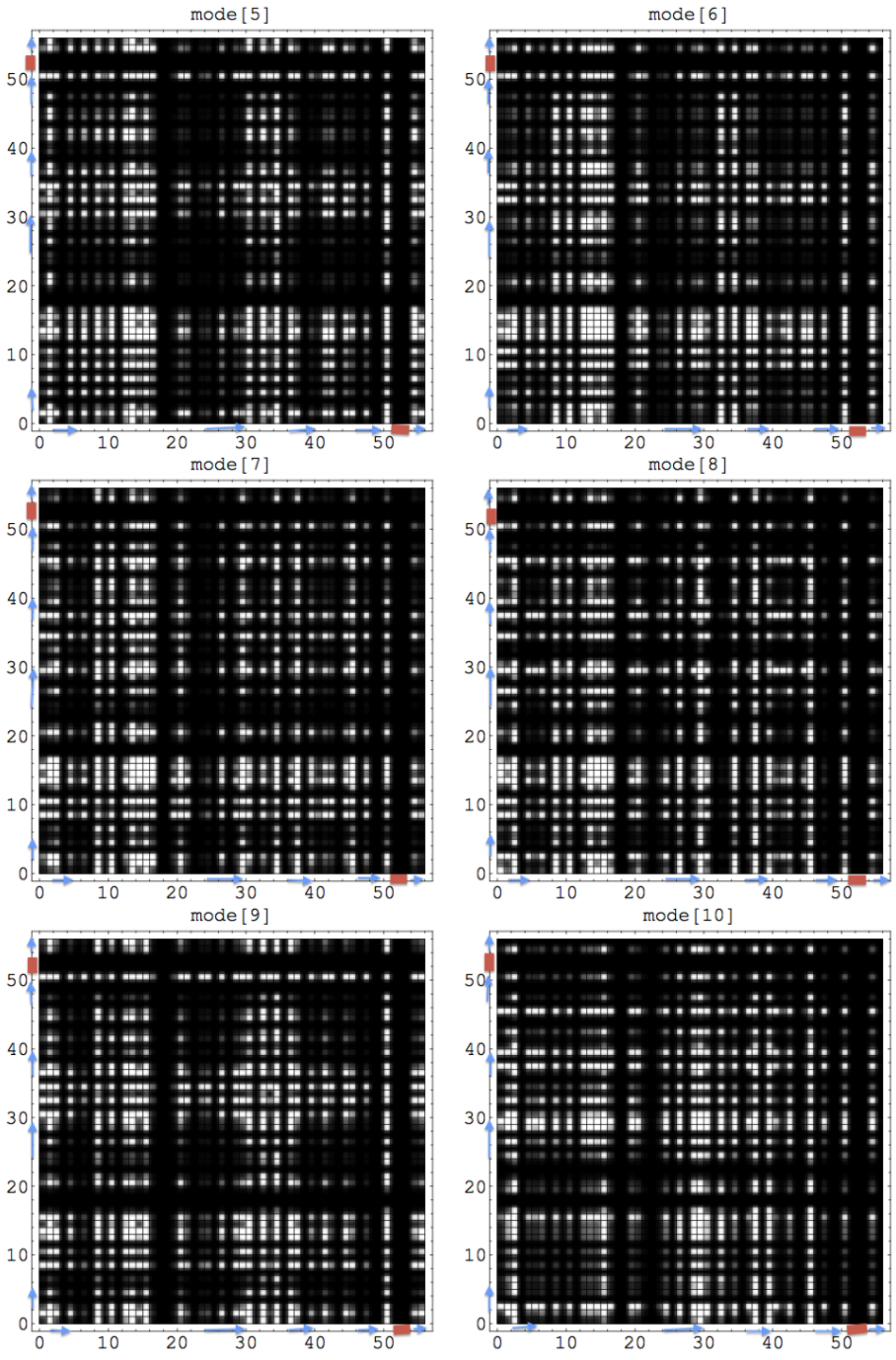


Figure 5.9 – Matrices représentant les modes propres de 5 à 10 tirés de l'analyse spectrale de la matrice de covariance totale avec alphabet réduit d'acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

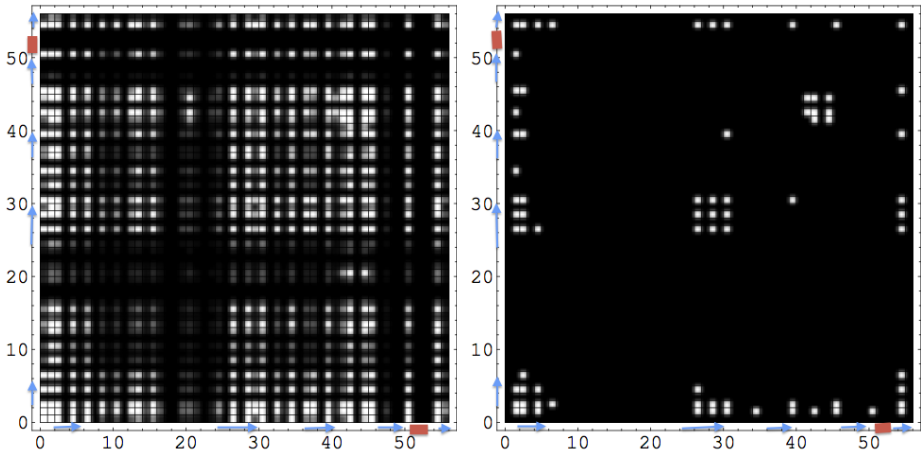


Figure 5.10 – Matrices correspondant aux modes propres 1 à 4 superposés et seuillés à 20 % tirés de l'analyse spectrale de la matrice de covariance avec alphabet réduit d'acides aminés. Les flèches bleues sur les axes représentent les feuillets beta de la protéine, et les rectangles rouges les hélices alpha.

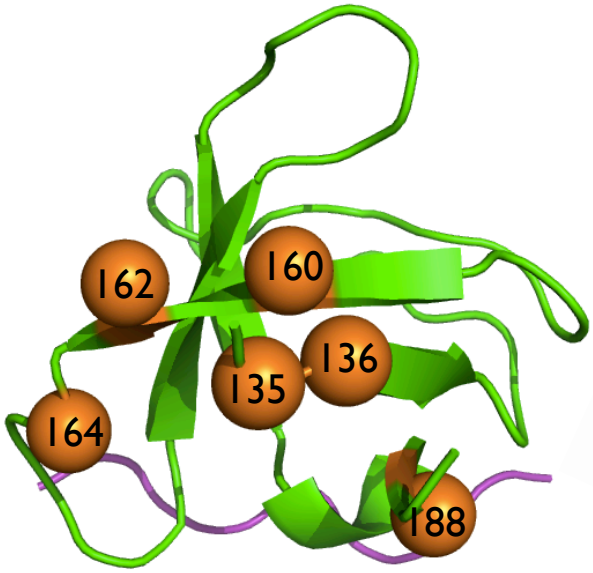


Figure 5.11 – Les positions 135-136-160-162-164-188 tirées de l'analyse spectrale de la matrice de covariance avec alphabet réduit d'acides aminés sont représentées par des sphères disposées sur le squelette peptidique de la protéine.

Motif	Nb séquences	Pourcentage
TKITRT	281	5.0 %
TLLTRT	200	4.0 %
LKITRT	119	2.4 %
KKITDT	116	2.3 %
RKITTT	97	1.9 %
TKITET	89	1.8 %
KKITTT	88	1.8 %
HKITQT	87	1.8 %
LLLTRT	84	1.7 %
TLLTRR	80	1.6 %
HKITET	80	1.6 %
Total	1321	25,9 %
Reste	3679	74,1 %

Table 5.6 – Liste des plus fréquents motifs sur les positions 135-136-158-160-164-188 (R158) avec un alphabet de 20 acides aminés.

5.5.1.4 Recherches d'homologues

Nous avons continué la suite du protocole, à savoir la recherche d'homologues, sur les deux ensembles de positions obtenus à partir des matrices de covariances avec l'alphabet complet et réduit d'acides aminés : 135-136-**158**-160-164-188, que nous noterons R158 et 135-136-160-**162**-164-188, notée R162.

Pour l'ensemble de positions R158 (alphabet complet), nous trions et regroupons les séquences suivant les motifs d'acides aminés les plus fréquents apparaissant sur ces 6 positions (tableau 5.6). Ainsi, nous obtenons 11 sous-ensembles de séquences qui représentent ensemble 26% des séquences de départ. Nous allons donc construire 12 profils de recherche BLAST correspondant à chacun des 11 sous-ensembles et au reste des séquences (74%). Nous trouvons 37 protéines homologues (non redondantes) à la protéine SH3-1CKA, sans aucun faux positif, alors que nous n'avons trouvé que 22 homologues en construisant un profil à partir de toutes les séquences (homologues en commun sur 2 tirages).

Motif	Nb séquences	Pourcentage
ARALRA	333	6.7 %
RRALAA	286	5.7 %
ALALRA	201	4.0 %
HRALDA	194	3.9 %
RLALAA	165	3.3 %
RRALDA	152	3.0 %
LRALRA	145	2.9 %
ALALRR	127	2.5 %
ARALRR	125	2.5 %
ARALDA	109	2.2 %
Total	1784	35,7 %
Reste	3163	64,3 %

Table 5.7 – Liste des plus fréquents motifs sur le motif 135-136-160-162-164-188 (R162) avec un alphabet réduit de 9 acides aminés.

De même pour l'ensemble de positions R162, nous regroupons les séquences suivant les motifs fréquents en utilisant toujours l'alphabet réduit d'acides aminés (tableau 5.7). Ainsi, nous obtenons 10 sous-ensembles de séquences qui représentent ensemble 36% des séquences de départ. Nous allons donc construire 11 profils de recherche BLAST correspondant à chacun des 10 sous-ensembles et au reste des séquences (64%). Les séquences utilisées pour les profils sont en revanche avec l'alphabet complet d'acides aminés. Nous trouvons 32 protéines homologues (non redondantes) à la protéine SH3-1CKA, sans aucun faux positif.

Toutes les protéines sont des domaines SH3, il n'y a pas de faux positifs.

Il est donc évident que le partitionnement en sous-ensembles moyenne moins l'information dans l'alignement de séquences, puisque nous considérons plusieurs sous-alignements, et permet de trouver plus d'homologues naturels.

On peut ajouter qu'on ne trouve pas exactement les mêmes homologues suivant le choix du motif. En effet, seuls 26 homologues sont communs aux tirages R158 et R162. Nous pourrions considérer toutes les positions de R158 et R162 pour regrouper en sous-ensembles les séquences, nous ajouterions alors encore plus d'information à chaque profil

NOMBRE HOMOLOGUES SH3	
Contexte	Nb homol. non red. (red.)
R158 (135-136-158-160-164-188)	37 (59)
R162 (135-136-160-162-164-188)	32 (61)
Toutes les séquences - tirage 1	19 (31)
Toutes les séquences - tirage 2	21 (33)
Tous confondus	49(70)
NOMBRE HOMOLOGUES SH3 EN COMMUN	
Contexte	Nb homol. non red. (red.)
Les 2 tirages sans groupe	22(34)
R158 et R162	26(50)
Commun à tous	16(29)

Table 5.8 – Pour la structure 1CKA du domaine SH3 : nombre d'homologues non redondants trouvés avec le protocole de recherche BLAST basé sur l'ensemble des séquences théoriques (sans groupes) et sur les ensembles de séquences construits sur les motifs. Entre parenthèses, le nombre d'homologues redondants.

Protéines homologues entre elles (>90%)					
G1	P46108	Q64010	Q63768	Q04929	P05433
G2	P32577	Q0VBZ0	P41241	P41240	P41239
G3	P62994	Q60631	Q07883	Q66II3	Q6GPJ9
G4	O75886	O88811	Q5XHY7	O93436	
G5	Q0U6X7	Q4WHP5	Q5BBL4		
G6	Q6WKZ7	Q5I0D6	Q8IVI9		
G7	P46109	P47941	Q5U2U2		
G8	Q62422	Q8MJ50	Q8MJ49		
G9	A4RF61	Q2GT05	Q7S6J4		
G10	Q6TGW5	Q6XJU9			
G11	P09769	Q02977			
G12	P70297	Q92783			
G13	P24604	P42680			
G14	P42686	P42690			

Table 5.9 – Pour la structure 1CKA du domaine SH3 : la liste des groupes de protéines homologues redondantes (90% de similarité) trouvées avec le protocole de recherche BLAST.

de recherche d'homologues. Néanmoins, rajouter trop de positions n'est pas forcément un bon choix. Si nous choisissons trop de positions, les sous-ensembles ne contiendront pas assez de séquences pour que les résultats soient améliorés.

Protéines homologues sorties sans groupe (redondants) :		
A0JNB0|FYN_BOVIN	P32577|CSK_RAT	Q28923|YES_CANFA
A1Y2K1|FYN_PIG	P34109|MYOD_DICDI	Q5PYH5|DLG1L_BRARE
O15259|NPHP1_HUMAN	P39688|FYN_MOUSE	Q5R4J7|GRB2_PONAB
O17972|NPHP1_CAEEL	P41239|CSK_CHICK	Q5U228|SH3B4_XENTR
O43295|SRGP2_HUMAN	P41240|CSK_HUMAN	Q5XHY7|STAM2_RAT
O43639|NCK2_HUMAN	P41241|CSK_MOUSE	Q60631|GRB2_MOUSE
O75044|FNBP2_HUMAN	P42680|TEC_HUMAN	Q62696|DLG1_RAT
O88811|STAM2_MOUSE	P42681|TXK_HUMAN	Q62844|FYN_RAT
P00523|SRC_CHICK	P42682|TXK_MOUSE	Q63768|CRK_RAT
P00525|SRC_AVISR	P42686|SRK1_SPOLA	Q64010|CRK_MOUSE
P00526|SRC_RSVP	P42690|SRK4_SPOLA	Q66II3|GRB2_XENTR
P00527|YES_AVISY	P46108|CRK_HUMAN	Q6GPJ9|GRB2B_XENLA
P05433|GAGC_AVISC	P46109|CRKL_HUMAN	Q7SDM3|MYO1_NEUCR
P06241|FYN_HUMAN	P47941|CRKL_MOUSE	Q812A2|SRGP2_MOUSE
P07947|YES_HUMAN	P62484|ABI2_MOUSE	Q8CBW3|ABI1_MOUSE
P09324|YES_CHICK	P63185|SRC_RSVSE	Q8TE67|ES8L3_HUMAN
P09769|FGR_HUMAN	P87378|CRK_XENLA	Q91WL0|ES8L3_MOUSE
P10936|YES_XENLA	P87379|GRB2A_XENLA	Q91Z67|FNBP2_MOUSE
P12931|SRC_HUMAN	Q02977|YRK_CHICK	Q921I6|SH3B4_MOUSE
P13115|SRC1_XENLA	Q04736|YES_MOUSE	Q9CX99|GRAP_MOUSE
P13116|SRC2_XENLA	Q04929|CRK_CHICK	Q9JJS5|SH3B4_RAT
P13406|FYN_XENLA	Q05876|FYN_CHICK	Q9NYB9|ABI2_HUMAN
P14084|SRC_AVISS	Q07014|LYN_RAT	Q9P0V3|SH3B4_HUMAN
P14085|SRC_AVIST	Q07883|GRB2_CHICK	Q9QY53|NPHP1_MOUSE
P14234|FGR_MOUSE	Q08012|DRK_DROME	Q9QZM5|ABI1_RAT
P15054|SRC_AVIS2	Q08881|ITK_HUMAN	Q9TU19|NPHP1_CANFA
P16333|NCK1_HUMAN	Q0VBZ0|CSK_BOVIN	Q9V9J3|SRC42_DROME
P17713|STK_HYDAT	Q12959|DLG1_HUMAN	Q9WUD9|SRC_RAT
P25020|SRC_RSVH1	Q13588|GRAP_HUMAN	Q9XYM0|CRK_DROME
P29355|SEM5_CAEEL	Q1LVQ2|SH3B4_BRARE	
P31007|DLG1_DROME	Q1LYG0|SPD2A_DANRE	

Protéines homologues sorties qu'avec les groupes (redondantes) :		
A7EK16|MYO1_SCLS1	P27447|YES_XIPHE	Q5XGP7|SKA2A_XENLA
O74653|POB1_SCHPO	P31693|SRC_RSVPA	Q63622|DLG2_RAT
O75962|TRIO_HUMAN	P42685|FRK_HUMAN	Q6DFH5|SH3Y1_XENLA
O89032|SPD2A_MOUSE	P43603|LSB3_YEAST	Q6PG29|SKAP2_BRARE
O93436|STAM2_CHICK	P98171|RHG04_HUMAN	Q7Z6B7|SRGP1_HUMAN
P07948|LYN_HUMAN	Q0KL02|TRIO_MOUSE	Q811D0|DLG1_MOUSE
P08630|BTKL_DROME	Q15700|DLG2_HUMAN	Q8AXQ3|SH3B4_SERQU
P24604|TEC_MOUSE	Q5FVW6|SKAP2_XENTR	Q91XM9|DLG2_MOUSE
P25911|LYN_MOUSE	Q5TCZ1|SPD2A_HUMAN	Q91Z69|SRGP1_MOUSE
P27446|FYN_XIPHE	Q5U597|SKA2B_XENLA	Q922K9|FRK_MOUSE

Protéines homologues sorties uniquement sans groupe (redondants) :
O17972|NPHP1_CAEEL

Liste de toutes les protéines homologues non redondantes					
A7EK16	G28	G59	P27446	P87378	Q8AXQ3
G1	G29	G68	P27447	P98171	Q8TE67
G16	G30	O15259	P29355	Q07014	Q91WL0
G18	G31	O17972	P31007	Q08012	Q922K9
G19	G32	O43639	P34109	Q08881	Q9QY53
G22	G38	O74653	P42680	Q1LYG0	Q9TU19
G23	G44	P08630	P42681	Q5XGP7	Q9V9J3
G24	G47	P14234	P42682	Q6DFH5	Q9XYM0
G25	G53	P16333	P42685	Q6PG29	
G26	G55	P17713	P43603	Q7SDM3	

Figure 5.12 – Liste des protéines homologues trouvées pour la protéine SH3-1CKA.

Structure (code pdb)	Nb acides aminés	Num. résidus	% identité
PDZ-1BE9	83	311-393	32,42%
PDZ-1QAU	112	14-125	30,79%
PDZ-2FE5	93	222-314	40,56%
SH3-1CKA	56	134-189	36,12%
SH3-1CSK	56	11-66	35,87%
SH3-1UTI	57	1-57	31,18%

Table 5.10 – Pour chaque structure des domaines PDZ et SH3 : nombre d'acides aminés considérés pour générer les séquences théoriques, les positions des résidus considérés (numérotation Pymol) et le pourcentage d'identité des alignements de séquences théoriques par rapport à la séquence native.

5.5.2 Analyse des covariances et recherches d'homologues pour d'autres domaines de protéines

5.5.2.1 Données

Nous disposons de six jeux de séquences théoriques générées à partir de six structures différentes, avec les paramètres utilisé dans le cas *old* détaillés dans le chapitre suivant. Nous avons choisi quelques structures de domaines PDZ et SH3. Dans le tableau 5.10, nous pouvons voir pour chaque structure, le pourcentage d'identité de l'ensemble des séquences théoriques par rapport à la séquence native. On peut remarquer que les séquences théoriques pour les domaines PDZ et SH3 sont identiques à 34% en moyenne par rapport à leur séquence naturelle correspondante.

5.5.2.2 Analyse des covariances et recherche d'homologues

Le détail des analyses des covariances pour les six structures PDZ et SH3 se trouvent en annexe. Les critères de sélection des résidus que nous utilisons sont heuristiques et intuitifs mais ils nous ont souvent donné des résultats cohérents en ce qui concerne la localisation spatiale des motifs sur les structures d'un même domaine. Dans les figures 5.13 et 5.14, on peut voir que les motifs sélectionnés pour chaque structure se recouvrent structurellement assez bien.

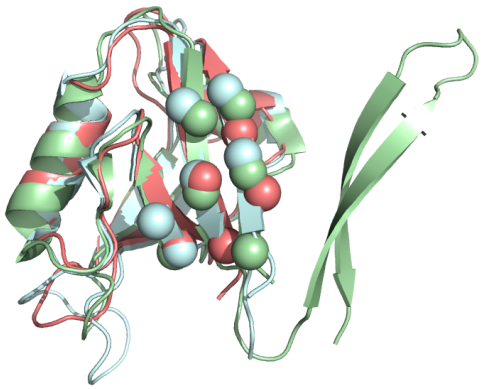


Figure 5.13 – Superposition des motifs des 3 structures des domaines PDZ : en vert 1QAU, en orange 1BE9 et en bleu 2FE5. (Pymol)

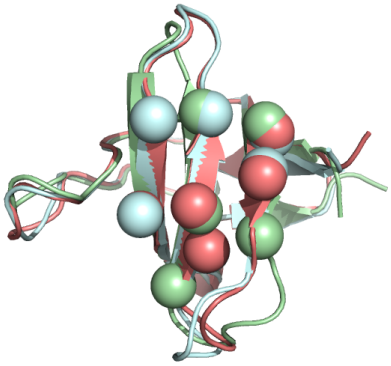


Figure 5.14 – Superposition des motifs des 3 structures des domaines SH3 : en vert 1UTI, en orange 1CKA et en bleu 1CSK. (Pymol)

Pour chacune des six structures, nous résumons dans le tableau 5.11 les résultats des étapes de calcul de covariance, d'analyse spectrale et de sélection du motif.

Structure	Nb modes	Seuil débruitage	% signal	Motif
PDZ-1BE9	10	25%	30%	312-314-358-389-391
PDZ-1QAU	5	25%	11%	15-17-19-63-94-96
PDZ-2FE5	3	18%	11%	226-228-277-279-306-308
SH3-1CKA	9	25%	48%	160-162-170-171-181
SH3-1CSK	7	38%	36%	37-39-48-51-57-59
SH3-1UTI	6	10%	43%	27-30-36-38-40

Table 5.11 – Pour chaque structure des domaines PDZ et SH3 : nombre de modes superposés et seuil de débruitage (pourcentage de la plus grande valeur des composants des vecteurs propres considérés), pourcentage du signal total des covariances correspondant au signal des modes, et positions du motif.

Struc.	Sans groupe			
PDB	t1	t2	commun	total
Domaine PDZ				
1BE9	101(0)/48	101(0)/48	101/48	48
1QAU	8(2)/3	7(1)/3	6/2	4
2FE5	83(0)/34	83(0)/34	83/34	34
Domaine SH3				
1CKA	90(0)/38	83(0)/35	82/33	40
1CSK	19(0)/6	19(0)/6	19/6	6
1UTI	56(0)/28	50(0)/26	50/26	28

Table 5.12 – Pour chaque structure des domaines PDZ et SH3 : nombre d'homologues trouvés avec un protocole basé sur l'ensemble des séquences théoriques (sans groupe) pour 2 tirages, commun et totals. Pour chaque cas, le premier nombre est le nombre d'homologues redondants trouvés, et le deuxième nombre est le nombre d'homologues non redondants. Entre parenthèses, le nombre de faux positifs soit le nombre d'homologues n'appartenant pas au bon domaine de protéines.

Struc.	Avec groupe				Avec et sans groupe	
PDB	t1	t2	commun	total	tous	commun
Domaine PDZ						
1BE9	107(0)/52	113(0)/54	100/51	55	120/56	96/23
1QAU	12(3)/5	13(3)/6	12/5	6	13/6	6/2
2FE5	88(0)/37	87(0)/36	87/36	37	88/39	83/15
Domaine SH3						
1CKA	118(0)/54	112(0)/50	110/50	54	121/58	82/18
1CSK	24(0)/8	24(0)/8	24/8	8	24/8	19/2
1UTI	72(0)/37	73(0)/36	69/35	38	76/38	50/9

Table 5.13 – Pour chaque structure des domaines PDZ et SH3 : nombre d'homologues trouvés avec un protocole basé sur les ensembles de séquences construits d'après les motifs (avec groupes) pour 2 tirages, en commun et totals. Pour chaque cas, le premier nombre est le nombre d'homologues redondants trouvés, et le deuxième nombre est le nombre d'homologues non redondants. Entre parenthèses, le nombre de faux positifs soit le nombre d'homologues n'appartenant pas au bon domaine de protéines.

5.6 Discussions

5.6.1 Utilisation de l'analyse spectrale

Nous nous sommes demandé si l'analyse spectrale avait réellement un effet positif sur le choix du motif. Nous avons alors tenté de travailler uniquement sur la matrice de covariance entière de la structure SH3-1CKA (matrices dans la figure 5.15). Nous pouvons constater que la diagonalisation de l'analyse spectrale supprime les valeurs fortes de la diagonale de la matrice. Ici, sans l'analyse spectrale, les valeurs les plus fortes qui conservées par le seuillage, font principalement partie de la diagonale. Nous obtenons alors le même motif avec 5 modes superposés et seuillés à 25 %. Néanmoins, d'un point de vue théorique, l'analyse spectrale permet de réduire le bruit dans les covariances.

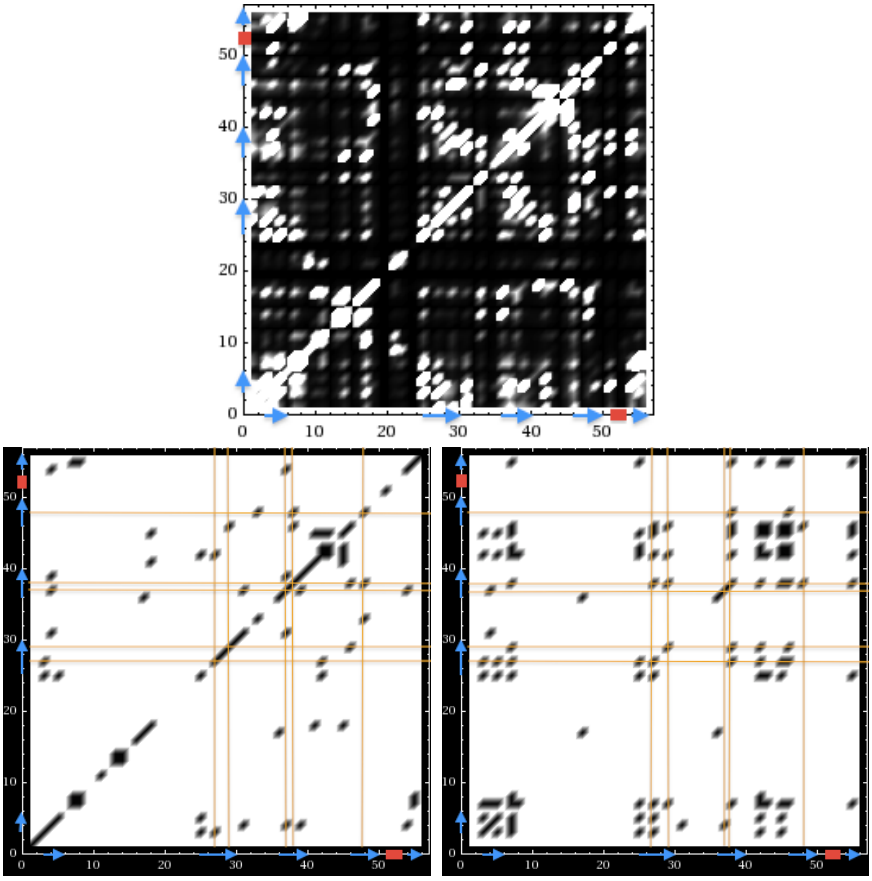


Figure 5.15 – Domaine SH3 de structure 1CKA. (En haut) Matrice de covariance entière. (En bas à gauche) Matrice de covariance entière seuillée à 10%. (En bas à droite) Matrice correspondant à la superposition des 9 premiers modes de l'analyse spectrale. Notation schématique des éléments de structures secondaires sur les axes des matrices : flèche bleue pour les feuillets β, rectangle rouge pour les hélices α.

5.6.2 Covariances sur des séquences naturelles

Nous avons tenté des calculs de covariance sur des alignements multiples obtenus à partir de séquences naturelles de la famille du domaine structurel PDZ. Nous disposons d'une MSA d'environ 1800 séquences, de 85 acides aminés. Nous nous sommes appuyés sur la séquence de la structure 2H3L pour construire l'alignement. Sur la matrice de covariance, on peut voir des zones plus fortement corrélées qui correspondent à l'hélice α_1 et au feuillet β_2 qui sont en contact avec le ligand. Les résultats sont assez bruités, et il est difficile d'en tirer des informations intéressantes.

Un autre problème s'est posé dans le cas des séquences naturelles. Il y a un pourcentage assez important de gap dans tout l'alignement. Les gaps sont pour l'instant considérés comme des acides aminés quelconques. Ainsi beaucoup de covariances impliquent des gaps, faussant tout résultat.

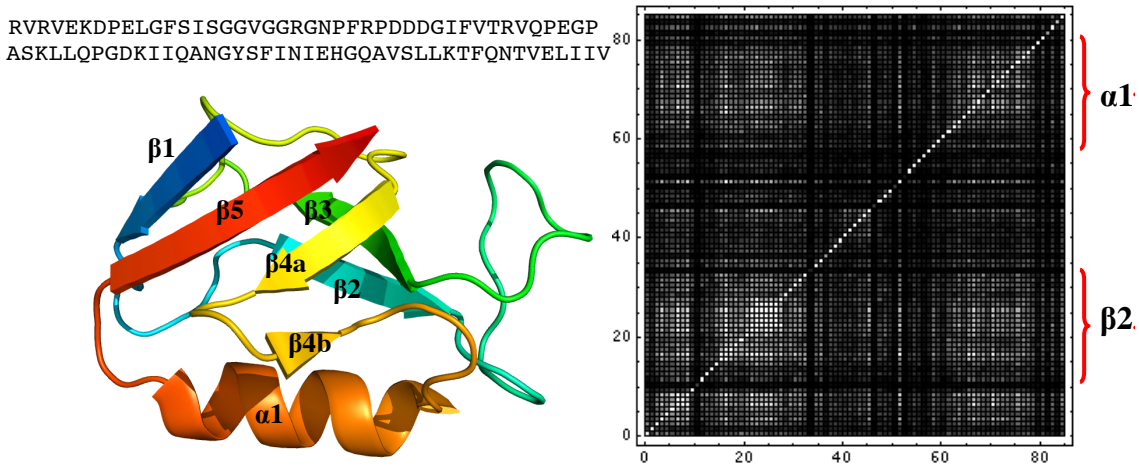


Figure 5.16 – (A gauche) Représentation Pymol du domaine PDZ de structure 2H3L et sa séquence. (A droite) Matrice de covariance entière.

5.7 Conclusion sur l'étude statistique de la génération de séquences théoriques par CPD

L'analyse des covariances nous a permis d'identifier les interactions entre acides aminés. La localisation spatiale des motifs trouvés pour des structures du même domaine, semble valider notre méthode d'un point de vue biologique, ou plus précisément structurel. Les positions des motifs sont généralement très proches dans la structure 3D, formant ainsi des réseaux de résidus qui stabiliseraient le rapprochement des structures secondaires.

La classification des acides aminés ne change pas trop l'aspect des matrices de covariance, elle gomme simplement les covariances entre acides aminés spécifiques et fait ressortir plus particulièrement les covariances entre types d'acides aminés. Il est effectivement plus intuitif et plus juste de se fier aux relations entre positions dues à leurs seules propriétés physico-chimiques.

Les requêtes Blast construisent à partir des séquences soumises, un profil statistique qui "moyenne" toutes ces séquences. En soumettant toutes les séquences théoriques d'une structure à Blast, on "moyenne" donc l'information sur les covariances. Ainsi en construisant des sous-ensembles de séquences, on les classe en fonction du motif, soit des covariances. L'information sur les covariances est donc un peu moins "moyennée" au sein de chaque groupe, ce qui doit améliorer la recherche d'homologues.

Grâce à l'analyse des covariances, les recherches Blast montrent que la construction et l'utilisation des groupes améliorent toujours et nettement le nombre d'homologues et que ces nouveaux homologues obtenus correspondent à juste quelques sous-ensembles de séquences correspondant au motif. Néanmoins, les résultats de recherche d'homologues sur les séquences théoriques sont bien moins bons que sur des séquences naturelles, de l'ordre de moins 88% en moyenne.

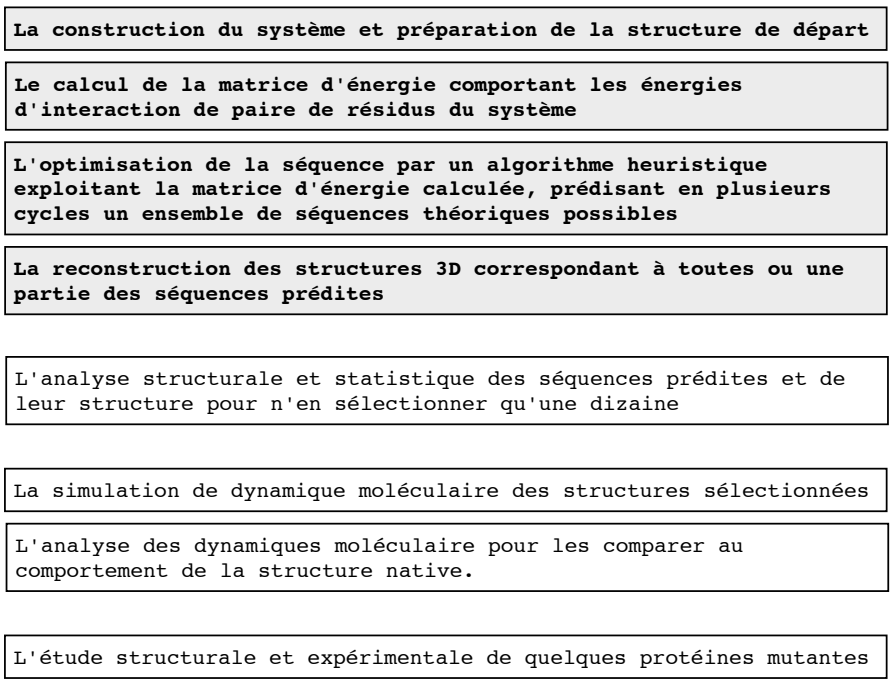
Nous avons montré que les performances de la recherche d'homologues à partir de nos séquences théoriques sont améliorées si on tient compte de l'information des covariances

entre différentes positions de la chaîne polypeptidique. Cela permet de retrouver une grande majorité des homologues naturels. L'utilisation combinée de séquences théoriques et expérimentales commence à être explorée par d'autres équipes et elle devrait permettre des performances supérieures. Une partie de ces résultats ont été décrite dans un article publié en 2010 (Schmidt am Busch, Sedano, Simonson, Plos One, 2010) (en annexe).

Chapitre 6

Génération de séquences théoriques

La suite de cette thèse se décompose en plusieurs étapes :



Dans ce chapitre, nous décrirons plus précisément les premières étapes de la génération de séquences (en gras ci-dessus) : l'algorithme de génération de séquences théoriques ainsi que les divers paramètres choisis (champ de force, modélisation du solvant, discrétisation du système, ...), et l'exploration de l'espace des séquences possibles et la reconstruction de leur structure. Ensuite, nous analyserons la qualité des séquences prédites suivant les protocoles choisis pour la génération.

Afin d'illustrer toutes les analyses réalisées dans cette thèse, nous nous appuierons sur des séquences théoriques générées à partir de la structure native du domaine SH3 de la protéine CRK-*mouse* de code PDB 1CKA. Cette structure semble être un bon système d'étude : elle est simple est bien caractérisée.

Le choix des meilleures séquences par la mise en place de filtres et les simulations de dynamique moléculaire seront décrits et analysés dans les chapitres suivants.

Nous explorons plusieurs jeux de paramètres pour la génération de séquences dans le but de réaliser un contrôle qualité de ces séquences prédites et de valider notre modèle. Cela nous permettra également de tester et améliorer nos protocoles de génération en utilisant des jeux de paramètres différents (champ de force, modèle électrostatique...), des énergies de référence optimisées avec différentes bases de données, en fixant différentes positions fonctionnelles ou spécifiques, ou en utilisant des modèles différents suivant la localisation spatiale des positions à prédire.

6.1 Génération de séquences théoriques

La première étape est donc la génération d'un ensemble de séquences théoriques, prédites à partir d'une structure 3D donnée. Aussi, il convient de décrire l'algorithme qui permet de calculer la matrice d'énergie, la fonction d'énergie et la représentation du solvant utilisés.

6.1.1 Calcul de la matrice d'énergie

La première étape consiste à calculer la matrice d'énergie qui contiendra les énergies d'interaction entre toutes les paires de résidus de la protéine, en autorisant successivement tous les types d'acides aminés possibles et tous les rotamères possibles. Nous maintenons fixés les atomes du squelette protéique (N, H, $C\alpha$, C et O) pendant toutes les étapes de

calculs, seuls les chaînes latérales sont mutées. Nous décrivons l'algorithme dans la figure 6.1.

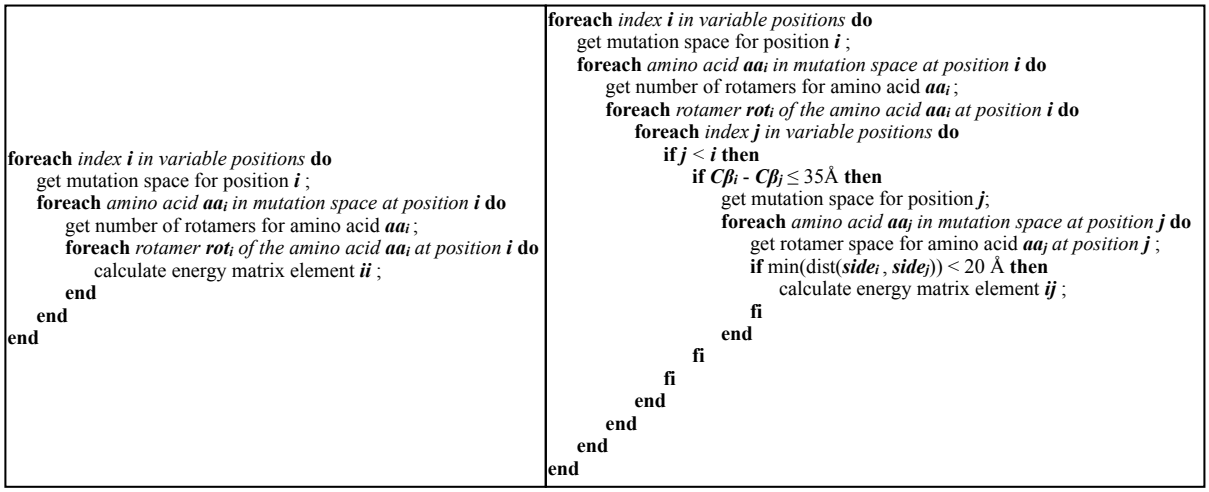


Figure 6.1 – A gauche, l'algorithme de calcul des termes diagonaux de la matrice d'énergie. A droite, l'algorithme de calcul des termes non-diagonaux de la matrice d'énergie.

Les calculs d'énergie sont réalisés avec le logiciel XPLOR [X-plor version 3.1 & NMR. 1992].

Les atomes des chaînes latérales sont construits géométriquement à partir des coordonnées des atomes du squelette, en se basant sur les paramètres des angles dièdres Φ et Ψ issus de la bibliothèque de rotamères Tuffery95 [Tuffery *et al.* 1991, 1997] (voir paragraphe sur la discrétisation de l'espace conformationnel).

6.1.1.1 Fonction d'énergie

L'énergie interne de la protéine est décrite par une fonction empirique issue du champ de force AMBER (ff99SB) [Weiner *et al.* 1984 ; Cornell *et al.* 1995 ; Wang *et al.* 2000 ; Hornak *et al.* 2006], reposant sur des critères de mécanique classique. Notre fonction d'énergie prend la forme :

$$E = E_{MM} + E_{Solv_{GBSA}} \tag{6.1}$$

que nous avons décrite dans la première partie de cette thèse.

Champ de force Pour rappel, le champ de force AMBER (*Assisted Model Building and Energy Refinement*) se définit comme une forme fonctionnelle sur des termes d'énergie et d'interaction, et un ensemble de paramètres attribués à un type d'atome, incluant aussi des améliorations portant sur les paramètres associés aux valeurs des angles dièdres Φ et Ψ.

Représentation implicite du solvant Le solvant est décrit de façon implicite par le modèle GBSA. Les surfaces accessibles au solvant (*Solvent accessible surface areas*) sont calculées avec l'algorithme de Lee & Richards [1971]. Cet algorithme utilise un rayon de "sonde" de 1,3 Å, les rayons de van der Waals et un paramètre de précision de 0,005. Différents ensembles de coefficients atomiques de solvatation σ_i peuvent être utilisés. Nous utilisons pour cette génération de séquences théoriques, un ensemble PHIA (pour *polar, hydrophobic, ionic, aromatic*) de coefficients atomiques : -0,005 pour les atomes non-polaires, -0,08 pour les atomes polaires, -0,04 pour les atomes aromatiques et $-0{,}10\,\mathrm{kcal/mol/\AA^2}$ pour les atomes ioniques.

6.1.2 Exploration des conformations

La deuxième étape de la méthode implique l'exploration d'un grand nombre de séquences. Le but est de trouver la combinaison séquences-rotamères qui minimise l'énergie globale. Pour cette étape, nous utilisons la procédure heuristique de Wernisch *et al.* [2000]. Lors d'un cycle heuristique, le squelette peptidique est conservé et reste fixe, puis une séquence d'acides aminés et leur rotamères sont choisis aléatoirement. Ensuite, pour une position i donnée, on minimise l'énergie totale de la séquence en sélectionnant le meilleur type d'acide aminé et son rotamère en gardant fixe le reste de la séquence. Ainsi, on itère le processus en $i+1$ et on boucle plusieurs fois sur la séquence jusqu'à obtenir une stabilité de l'énergie totale. On obtient alors une séquence théorique. Pour chaque protéine, nous répétons ce cycle 200 000 fois généralement, en choisissant une nouvelle séquence aléa-

toire à chaque cycle. Les Glycines, Prolines et Cystéines sont "gelées". Ces acides aminés peuvent avoir des effets sur le repliement de la protéine difficiles à prendre en compte par notre méthode. Leur mutation pourrait donc avoir des conséquences notables sur la structure de la protéine. Par conséquent, ces résidus sont fixés en séquence et en structure dans l'état natif.

L'avantage de cet algorithme est qu'en multipliant les points de départ aléatoires, il permet en un temps réduit, d'explorer un espace des conformations et des séquences plus important qu'avec une procédure déterministe. Pour chacun des cycles heuristiques, la meilleure séquence est retenue, conduisant ainsi à un total de 200 000 séquences. Un premier filtre éliminant les séquences redondantes est alors appliqué. Puis les séquences peuvent être classées selon leur score énergétique. La procédure est implémentée dans un programme C/C++ appelée PROTEUS.

6.1.2.1 Discrétisation de l'espace conformationnel

Pour une structure donnée, la séquence d'acides aminés la plus favorable correspond à celle qui maximisera la différence d'énergie libre entre l'état replié et l'état déplié. Pour identifier cette séquence, nous modélisons ces deux états. Pour réduire la complexité de l'espace conformationnel des chaînes principales et latérales, nous le discrétisons. Même si des algorithmes plus réalistes de design multi-états avec un ensemble de squelettes protéiques existent [Su & Mayo 1997 ; Allen & Mayo 2012], nous avons fait le choix de considérer la chaîne principale comme fixe.

Nous décrivons l'espace conformationnel des chaînes latérales par les angles de torsions qui décrivent la rotation des groupes chimiques autour des liaisons. Les conformations préférentielles constituent les "rotamères" cataloguées dans des librairies [Janin & Wodak 1978 ; Ponder & Richards 1987 ; Kono & Doi 1994 ; Lovell *et al.* 1999 ; McGregor *et al.* 1987 ; Dunbrack & Karplus 1993b ; Peterson *et al.* 2004 ; Xiang & Honig 2001]. Nous utilisons ici la bibliothèque de rotamères Tuffery95. Cette bibliothèque contient un total

de 219 rotamères répertoriés dans le tableau 6.1, dont certains sont illustrés en exemple dans la figure 6.2.

Acide aminé	Nb rotamères	Acide aminé	Nb rotamères
ALA	1	ILE	7
ASP	5	LEU	9
ASN	11	LYS	48
ARG	39	MET	17
CYS	3	PHE	4
GLU	12	SER	3
GLN	19	TYR	8
HIS	9	THR	3
HSP	9	TRP	8
VAL	3		

Table 6.1 – Nombre de rotamères pour chaque acide aminé dans la bibliothèque de Tuffery. HSP représente l'Histidine doublement protonnée sur les atomes ND1 et ND2 et le HIS est simplement protonné sur ND1.

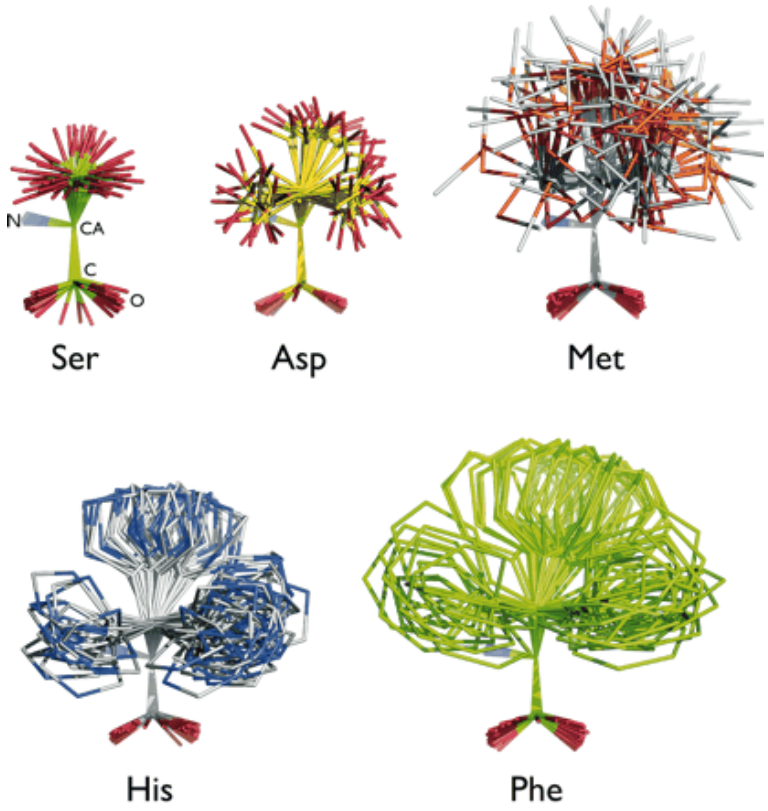


Figure 6.2 – Exemples de rotamères sur plusieurs acides aminés.

6.1.2.2 États dépliés et repliés, et énergie de repliement

Puisqu'aucune méthode actuelle n'est en mesure d'apporter une information structurale pour l'état déplié avec un même niveau de détail que pour l'état natif, nous utilisons un modèle assez simple. Il décrit l'état déplié par une chaîne polypeptidique étendue, où les chaînes latérales des acides aminés interagissent essentiellement avec le solvant et les groupes voisins du squelette peptidique. Comme les chaînes latérales n'engagent que très peu d'interactions entre elles, ce modèle considère que l'énergie libre de l'état déplié est uniquement dépendante de la composition en acides aminés et non de la séquence. L'approche la plus courante consiste à déterminer des énergies de référence pour chaque type d'acide aminé, représentant ainsi sa contribution individuelle à l'énergie libre de l'état déplié. Nous considérons en fait une collection de tripeptides Ala-X-Ala, où X est l'acide aminé à une position donnée. [Dahiyat & Mayo 1996]

L'énergie de repliement est définie par :

$$\Delta E = E_{folded} - E_{unfolded} \tag{6.2}$$

où $E_{unfolded} = \sum_i E_{ref_i}$.

Nous optimisons empiriquement les énergies de référence E_{ref_i} pour chaque type d'acide aminé i, de façon à obtenir des compositions d'acides aminés correspondant à celles dans les petits ensembles PFAM correspondant aux familles des domaines SH2, SH3 et PDZ [Schmidt am Busch *et al.* 2008b].

6.1.3 Reconstruction des structures 3D

A partir de chaque séquence explorée par PROTEUS, un structure 3D peut être reconstruite. En effet, les coordonnées du squelette peptidique sont maintenues fixées durant toute la procédure. Il suffit donc de positionner chaque acide aminé dans le rotamère

prédit. On réalise ensuite une minimisation de 200 pas afin de réajuster l'ensemble des chaînes latérales et ainsi éliminer les conflits stériques.

6.2 Différents scénarios de génération de séquences théoriques

Les paramètres utilisés dans les différentes générations sont listés dans le tableau 6.2 et les énergies de références pour chaque acide aminé dans le tableau 6.3. Les différents scénarios de génération de séquences sont détaillés juste après.

Génération	Champ de	Biblio.	GB ?	Détails				Correc. surf.		Minimisation		Filtres	
séquences	force	rota.		ε	α	σ_i	cutoff	buriedfactor	threshold	ij	i	$C_\beta - C_\beta$	\$dmin
Gen1	CHARMM	Tuff95	non	16	-	MF	-	-	-	-	-	-	-
Gen2	AMBER	Tuff95	oui	16	1	PHIA	999	0,65	0,3	15	15	35Å	20Å
Gen3	AMBER	Tuff95	oui	4	1	unif/-0,03	999	0,65	0,3	15	15	35Å	20Å
Gen4	AMBER	Tuff95	oui	16	1	PHIA	999	0,65	0,3	15	15	35Å	20Å

Table 6.2 – Liste des paramètres pour la génération de séquences théoriques. Pour les champs de force, CHARMM = toph19 et AMBER = ff99SB.

Pour représenter un modèle de l'état déplié cohérent, il est nécessaire d'ajouter une correction empirique aux énergies de référence. Pour chaque acide aminé I, une correction e_x est définie, où x représente le type d'acide aminé courant à la position I. Il y a 17 valeurs à optimiser, pour les 17 types d'acides aminés que l'on peut muter. Nous optimisons la valeur e_x jusqu'à obtenir une composition "cible" en acides aminés pour l'ensemble des séquences prédites pour des protéines tests. La composition "cible" en acides aminés est la moyenne des fréquences en acides aminés f_x^{exp} de l'ensemble des séquences sauvages des petits alignements PFAM des familles SH2, SH3 et PDZ. Nous procédons ensuite par itération, en partant initialement avec $e_x = 0$. A chaque itération, et pour chaque protéine test, 30000 séquences sont calculées. Les fréquences des acides aminés dans ces séquences prédites, f_x^{calc}, moyennées pour toutes les séquences, positions, et protéines tests, sont

comparées aux fréquences cibles f_x^{exp}. La correction e_x est alors modifiée en respectant la relation de Boltzmann :

$$e_x^{new} = e_x^{old} + 0.5ln\frac{f_x^{exp}}{f_x^{calc}} \tag{6.3}$$

Acide aminé	Gen1	Gen2-3-4	Acide aminé	Gen1	Gen2-3-4
ALA (A)	-11,307	-10.052	LEU (L)	-12,600	-12.305
ARG (R)	-25,043	-22.278	LYS (K)	-22,214	-21.628
ASN (N)	-17,180	-17.438	MET (M)	-13,922	-12.557
ASP (D)	-19,826	-20.713	PHE (F)	-17,412	-18.246
CYS (C)	0,000	0.000	PRO (P)	0,000	0.000
GLN (Q)	-17,940	-18.439	SER (S)	-13,450	-12.812
GLU (E)	-21,257	-20.787	THR (T)	-12,583	-11.840
GLY (G)	0,000	0.000	TRP (W)	-20,983	-18.881
HIS (H)	-20,389	-18.269	TYR (Y)	-20,274	-20.988
ILE (I)	-12,320	-10.255	VAL (V)	-11,481	-9.743

Table 6.3 – Énergies de référence (kcal/mol) pour chaque acide aminé dans son état déplié pour la génération Gen1-*old*, et pour les générations Gen2-LIG, Gen2-CORE, Gen3 et Gen4.

6.2.1 Gen1 : travaux antérieurs

Nous avons voulu comparer les comportements des dynamiques des protéines mutantes choisies avec celles d'une protéine mutante générée par notre équipe il y a quelque temps et sur laquelle j'ai réalisé un certain nombre d'études expérimentales en début de thèse (voir chapitre sur l'étude expérimentale). Cette séquence théorique (que nous noterons Gen1-*old*) a été générée à partir du même domaine SH3 c-Crk de structure 1CKA, avec des paramètres un peu différents. Schmidt am Busch *et al.* [2008a] avaient choisi d'utiliser le champ de force CHARMM19, le modèle CASA, les coefficients surfaciques MF (pour *Modified Fraternali*) [Fraternali & van Gunsteren 1996] au lieu des paramètres PHIA, le coefficient $\epsilon = 16$, et seuls les résidus Glycine, Proline, Cystéine ont été fixés.

Les énergies de référence (voir tableau 6.3) sont optimisées en générant des séquences sur les domaines SH3 : 1GCQ, 1CKA et 1SHG.

Cette génération de séquences théoriques nous servira de point de comparaison pour la plupart de nos analyses, et nous la nommerons Gen1-*old*.

Type atome	MF	PHIA
apolaire	0,0119	-0,005
aromatique	0,0119	-0,04
polaire	-0,0597	-0,08
ionique	-0,15	-0,10

Table 6.4 – Paramètres de solvatation atomique (kcal/mol/Å^2) des coefficients surfaciques MF et PHIA pour différents types d'atomes.

6.2.2 Gen2 : travaux actuels

Nous avons généré deux ensembles différents de séquences théoriques à partir de la structure 1CKA. Après le calcul d'une matrice d'énergie commune, dans un cas on fixe les positions du cœur hydrophobe, dans l'autre les positions fonctionnelles en contact avec le ligand. Nous nommerons ces générations Gen2-CORE et Gen2-LIG.

Les énergies de référence (voir tableau 6.3) sont optimisées en générant des séquences sur les domaines SH3 : 1ABO, 1CKA et 2PTK.

6.2.2.1 Choix des positions fonctionnelles

Nous sélectionnons 6 acides aminés à partir d'alignements structuraux de plusieurs domaines SH3 et en étudiant visuellement les contacts avec le ligand PPPALPPKKR, comme on peut le voir dans la figure 6.3. Cependant, nous faisons le choix de ne considérer que 3 résidus fonctionnels "gelés" pour générer des séquences théoriques : PHE141, TRP169 et TYR186.

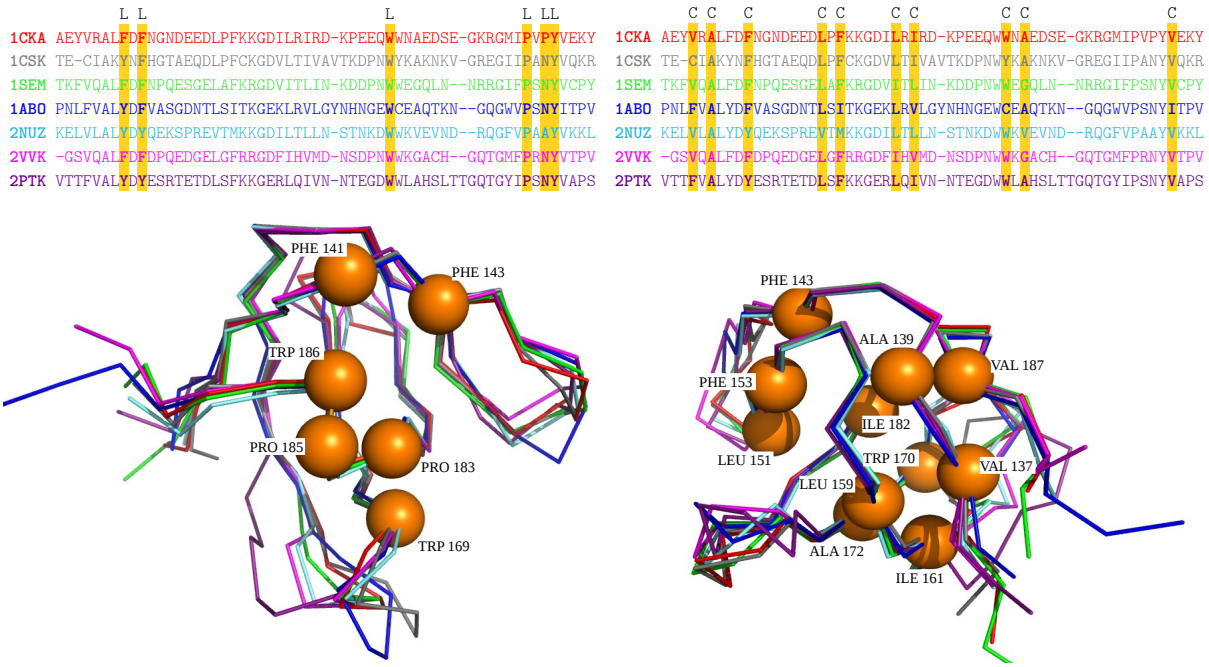


Figure 6.3 – A gauche : alignement de séquences de 7 domaines SH3. Les positions considérées fonctionnelles (en contact avec le ligand) sont marquées d'un *L*. Les structures de ces domaines sont également alignées (représentation *ribbon* sous Pymol), et les 6 positions fonctionnelles sont représentées par des sphères. A droite : alignement de séquences de 7 domaines SH3. Les positions considérées comme représentatives du cœur hydrophobe sont marquées d'un *C*. Les structures de ces domaines sont également alignées (représentation *ribbon* sous Pymol), et les 11 positions du cœur sont représentées par des sphères.

6.2.2.2 Choix des positions du cœur hydrophobe

Les interactions hydrophobes sont des facteurs importants dans le repliement et la stabilité des structures protéiques. Bien que les biologistes fassent souvent référence à l'appartenance de tel ou tel résidu au cœur hydrophobe d'une protéine pour expliquer les propriétés de ce résidu, il n'existe pas de définition unanime du cœur hydrophobe. Certaines définitions prennent en compte la conservation au cours de l'évolution des résidus hydrophobes "enfouis" tandis que d'autres ne s'appuient pas sur une analyse séquentielle [Hirakawa *et al.* 1999]. Quelques algorithmes ont ainsi été mis en place pour définir le cœur hydrophobe des protéines de manière systématique. Parmi eux, un algorithme proposé par Swindells [1995] décrit le cœur hydrophobe comme la collection des résidus possédant une accessibilité faible au solvant, appartenant à des régions de structures se-

condaires régulières et dont les chaînes latérales non polaires interagissent en partie entre elles. En général, ces trois propriétés sont en effet utilisées par les expérimentateurs pour définir le cœur hydrophobe de façon empirique.

Les résidus appartenant au cœur hydrophobe d'une protéine sont donc des résidus conservés dans les familles structurales et jouant un rôle important pour la stabilité de la protéine tant au moment de son repliement que dans sa structure native.

Nous sélectionnons donc 11 acides aminés à partir d'alignements structuraux de plusieurs domaines SH3 pour définir le cœur hydrophobe (voir figure 6.3) : VAL137, ALA139, PHE143, LEU151, PHE153, LEU159, ILE161, TRP170, ALA172, ILE182 et VAL187. Ainsi nous mutons toutes les positions de la protéine, exceptés ces 11 résidus conservés dans leur type natif dont nous optimisons uniquement leur rotamère. Nous mutons donc principalement la surface de la protéine en conservant les contraintes induites par le cœur hydrophobe.

6.2.3 Gen3 : d'autres paramètres de génération

Ce nouvel ensemble de générations de séquences théoriques sera appelé Gen3.

6.2.3.1 Hypothèses

Dans cette thèse, jusqu'à présent, nous utilisions pour la génération de séquences théoriques le modèle GB (*Generalized-Born*) et l'ensemble des coefficients surfaciques PHIA (coefficients σ_i). D'autres générations ont été réalisées dans notre équipe avec un coefficient surfacique uniforme pour tous les atomes, de $-0{,}03\,\mathrm{kcal/mol/\AA^2}$. Le meilleur protocole était une génération avec ϵ très grand et $\sigma = \mathrm{PHIA}$. En effet, plus on augmentait la valeur de ϵ (4, 10, 16, 24, 32), plus les résultats s'amélioraient. Avec une constante ϵ grande, nous avions l'impression que le terme électrostatique n'apportait pas une grande contribution lorsqu'il était combiné aux coefficients surfaciques PHIA.

Nous avons pu constater que les termes du modèle GB jouent probablement un rôle important pour empêcher l'enfouissement des résidus ioniques, mais c'est très nettement le terme PHIA pour les coefficients surfaciques qui donnent *in fine* le bon équilibre des résidus hydrophobes dans le cœur, aidé par le terme van der Waals. Dans le cas avec un coefficient surfacique uniforme (-0,03), le modèle ne peut compter que sur les termes GB pour faire la ségrégation hydrophobe, et cela ne semble pas suffisant.

6.2.3.2 Protocole

L'idée ici, est dans un premier temps, de générer avec un coefficient surfacique uniforme (non-PHIA) des séquences théoriques dont on aurait fixé le cœur hydrophobe dans le type natif. Ainsi, nous pourrions peut-être améliorer la prédiction des réseaux de charges en surface de la protéine. Dans un deuxième temps, nous générerons des séquences théoriques avec toujours le coefficient surfacique uniforme (de -0,03) mais en fixant les positions d'un cœur hydrophobe mutant sélectionné depuis une génération avec $\epsilon = 16$ et $\sigma =$ PHIA. Ainsi, le cœur hydrophobe serait prédit avec un bon équilibre des résidus hydrophobes, puis la surface serait mieux prédite sans les coefficients surfaciques PHIA.

Pour ces nouvelles générations de séquences théoriques, nous avons considéré trois cœurs hydrophobes différents que nous avons fixés : le cœur hydrophobe natif, et deux cœurs hydrophobes mutants provenant de deux séquences mutantes choisies pour la simulation en dynamique, à savoir 1CKA-LIG-V99 et 1CKA-LIG-W105 (voir détails dans le chapitre suivant). Ces deux séquences mutantes font parties de l'ensemble de séquences théoriques générées avec les positions fonctionnelles fixes détaillé précédemment dans ce chapitre.

6.2.3.3 Protocole 1 : cœur hydrophobe fixé de 11 résidus

Données Nous avons sélectionné 11 positions pour définir le cœur hydrophobe dans le protocole 1. Ce choix a déjà été détaillé précédemment dans ce chapitre. Nous pouvons

voir en rouge dans la figure 6.4 les contacts avec le solvant des positions choisies. Le tableau 6.5 récapitule les compositions résidus des cœurs hydrophobes natif et mutants.

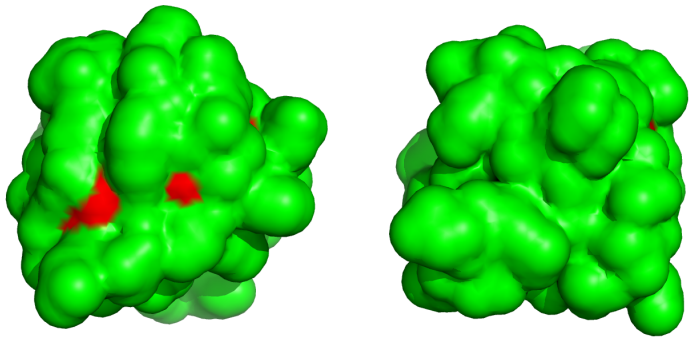


Figure 6.4 – Surface de la structure 1CKA : en rouge les 11 positions du cœur hydrophobe choisies. Vue de droite obtenue à partir de celle de gauche par rotation de 180 ° autour d'un axe vertical.

Protéine	137	139	143	151	153	159	161	170	172	182	187
WT	VAL	ALA	PHE	LEU	PHE	LEU	ILE	TRP	ALA	ILE	VAL
LIG-V99	VAL	**CYS**	PHE	LEU	PHE	LEU	ILE	**VAL**	ALA	ILE	VAL
LIG-W105	VAL	ALA	PHE	LEU	**HIS**	**MET**	ILE	TRP	ALA	ILE	VAL

Table 6.5 – Composition en acides aminés du cœur hydrophobe à 11 positions pour la protéine sauvage 1CKA-WT, et les protéines mutantes 1CKA-LIG-V99 et 1CKA-LIG-W105.

6.2.3.4 Protocole 2 : cœur hydrophobe de 19 résidus

Données Nous avons donc sélectionné 19 positions pour définir le cœur hydrophobe dans le protocole 2. Cette sélection a été obtenu avec un seuil de 40% représentant le pourcentage de surface du résidu en contact avec le solvant pour le considérer comme enfoui. Nous pouvons voir en rouge dans la figure 6.5 les contacts avec le solvant des positions choisies. Ce cœur hydrophobe est bien évidemment plus exposé au solvant que celui à 11 résidus. Néanmoins, pour mesurer les bénéfices ou inconvénients des nouveaux paramètres sur la prédiction de la surface, il convenait de tester plusieurs partitionnements et définitions de la surface et du cœur. Le tableau 6.6 récapitule les compositions résidus des cœurs hydrophobes natif et mutants.

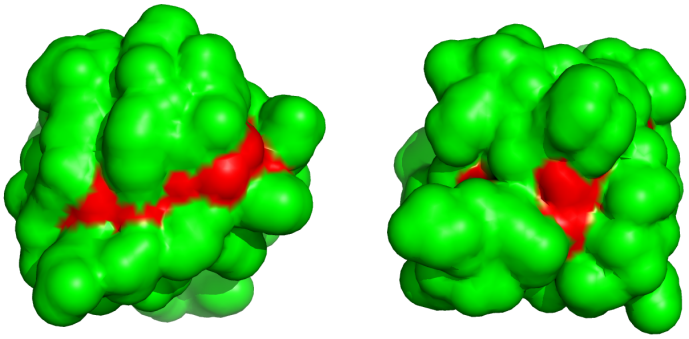


Figure 6.5 – Surface de la structure 1CKA : en rouge les 19 positions du cœur hydrophobe choisies. Vue de droite obtenue à partir de celle de gauche par rotation de 180 ° autour d'un axe vertical.

Protéine	137	139	140	143	150	151	153	157	158	159
WT	VAL	ALA	LEU	PHE	ASP	LEU	PHE	ASP	ILE	LEU
LIG-V99	VAL	**CYS**	LEU	PHE	**MET**	LEU	PHE	**LEU**	**LEU**	LEU
LIG-W105	VAL	ALA	LEU	PHE	**MET**	LEU	**HIS**	**LEU**	**LEU**	LEU

Protéine	161	170	171	172	174	181	182	184	187	-
WT	ILE	TRP	ASN	ALA	ASP	MET	ILE	VAL	VAL	-
LIG-V99	ILE	**VAL**	**GLU**	ALA	**SER**	**LEU**	ILE	**ASP**	VAL	-
LIG-W105	ILE	TRP	**VAL**	ALA	**SER**	**LEU**	ILE	**LEU**	VAL	-

Table 6.6 – Composition en acides aminés du cœur hydrophobe à 19 positions pour la protéine sauvage 1CKA-WT, et les protéines mutantes 1CKA-LIG-V99 et 1CKA-LIG-W105.

6.2.4 Gen4 : présence du ligand

Pour cette génération, notée "Gen4", seules les Prolines, Cystéines et Glycines de la séquences sauvages de la protéine 1CKA ne sont pas mutées. Ici, la présence du peptide PRO-PRO-PRO-ALA-LEU-PRO-PRO-LYS-LYS-ARG pendant le calcul de la matrice d'énergie contraint directement la mutation des acides aminés proches du ligand. Nous pouvons voir dans la figure 6.6 la position du ligand par rapport à la protéine SH3-1CKA.

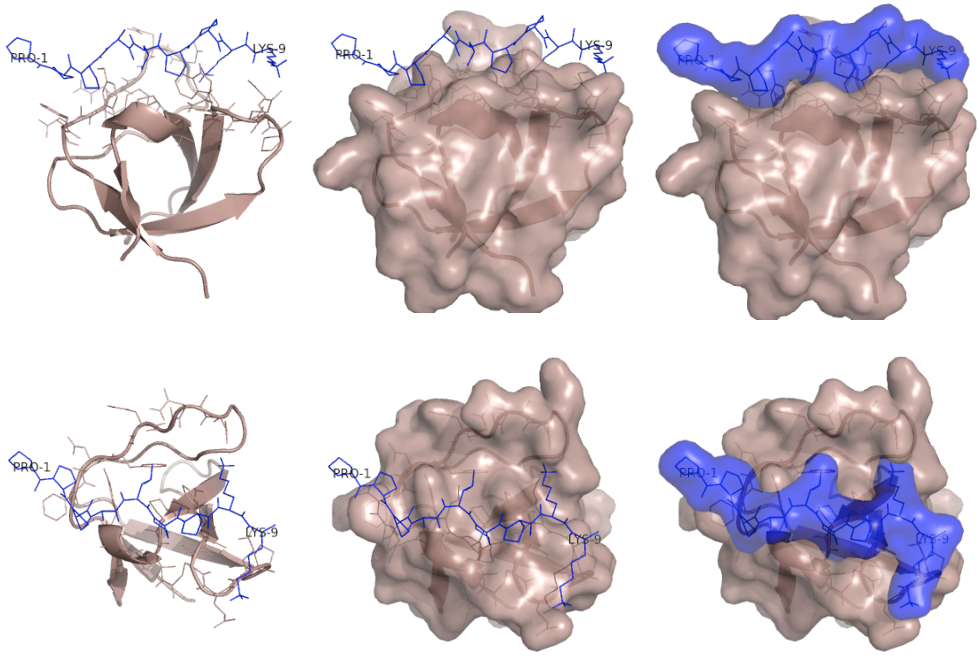


Figure 6.6 – Différentes représentations Pymol de la protéine 1CKA et son ligand (avec les chaînes latérales ou avec la surface) (Pymol).

6.3 Caractérisation des ensembles de séquences théoriques

Nous considérons ici les 5000 séquences théoriques de meilleure énergie PROTEUS, afin de juger de la qualité des ensembles de séquences prédites.

Dans le tableau 6.7, nous pouvons observer que les générations avec les positions du cœur hydrophobe natif fixées ont les meilleurs scores de similarité par rapport à l'alignement de séquences naturelles PFAM. En particulier les générations Gen2-CORE et Gen3-coreWT-19aas. Rien d'étonnant à cela puisque ces positions sont assez bien conservées dans l'alignement PFAM. Il en est de même pour le score de similarité par rapport à la séquence sauvage.

Par contre, il y a une différence notable entre les scores de similarité contre l'alignement PFAM pour les générations Gen3-coreWT-11aas et Gen2-CORE. On passe de 48 à 29 alors que les mêmes positions du cœur hydrophobe natif ont été fixées. Finalement, l'hypothèse que l'utilisation d'un coefficient surfacique uniforme (non PHIA) améliorerait la prédiction des réseaux de charges en surface semble être discutable. En effet, il y a moins de mutations ioniques par rapport à la séquences sauvage pour la génération Gen3-coreWT-11aas que dans la génération Gen2-CORE. Les scores d'identité et de similarité par rapport à la séquence sauvage sont améliorés également. Néanmoins, nous introduisons plus de mutations radicales et de charge par rapport à la séquence sauvage. Et les séquences de la génération Gen3-coreWT-11aas sont plus éloignées des séquences naturelles de l'alignement PFAM que les séquences de la génération Gen2-CORE.

En revanche, la génération Gen4 semble être le bon compromis. En effet, ses scores de similarité avec la séquence sauvage et avec l'alignement PFAM font partie des meilleures. La présence du ligand lors de la génération de séquences théoriques ajoute des contraintes plus fortes et plus nombreuses que le simple gel des positions fonctionnelles choisies,

comme pour la génération Gen2-LIG dont la qualité des séquences semblent moins bonne. Nous n'avons peut-être pas considéré suffisamment de positions en contact avec le ligand.

Les générations Gen2 et Gen3 ont des points isoélectriques moyens autour de 9, alors celui de la séquence sauvage est de 4,5 et celui de l'alignement PFAM est de 5,3 en moyenne. Ces valeurs sont influencées directement par l'abondance des acides aminés, et donc par les énergies de références. La génération Gen1 respecte un peu mieux les fréquences d'apparition des acides aminés : les énergies de références étaient peut-être plus adaptées.

Ensemble de séquences	CONTRE 1CKA-WT					pI	CONTRE PFAM	
	Ident.	Simil	% mut. radic.	% mut. de charge	% mut. ioniques	th. moy.	Simil.	% mut. radic.
PFAM	33.60	92.48	19%	40%	20%	5.27	67.95	25%
WT	100.00	307.00	0%	0%	0%	4.56	75.80	21%
Gen1-*old*	37.06	120.30	17%	36%	14%	6.13	38.48	32%
Gen2-LIG	38.25	101.19	17%	41%	18%	8.74	30.79	31%
Gen2-CORE	39.58	113.42	18%	38%	18%	8.89	48.14	29%
Gen3-coreWT-11aas	43.23	115.11	20%	43%	12%	9.16	29.69	32%
Gen3-coreWT-19aas	54.48	154.85	19%	34%	7%	9.01	41.57	29%
Gen3-core1-11aas	39.64	98.34	18%	42%	12%	9.01	17.04	31%
Gen3-core1-19aas	39.93	99.31	18%	37%	12%	9.01	16.24	33%
Gen3-core2-11aas	40.10	107.89	22%	45%	14%	9.01	25.29	33%
Gen3-core2-19aas	39.27	105.80	22%	39%	13%	9.01	19.98	36%
Gen4	41.71	117.27	15%	37%	12%	7.33	46.02	25%

Table 6.7 – Caractérisation des ensembles de séquences théoriques générées à partir de la structure SH3-1CKA. Moyenne des scores d'identité, de similarité, moyenne des pourcentages de mutations radicales, de mutations de charge, de mutations ioniques contre la séquence sauvage et contre l'alignement PFAM des domaines SH3. Point isoélectrique moyen.

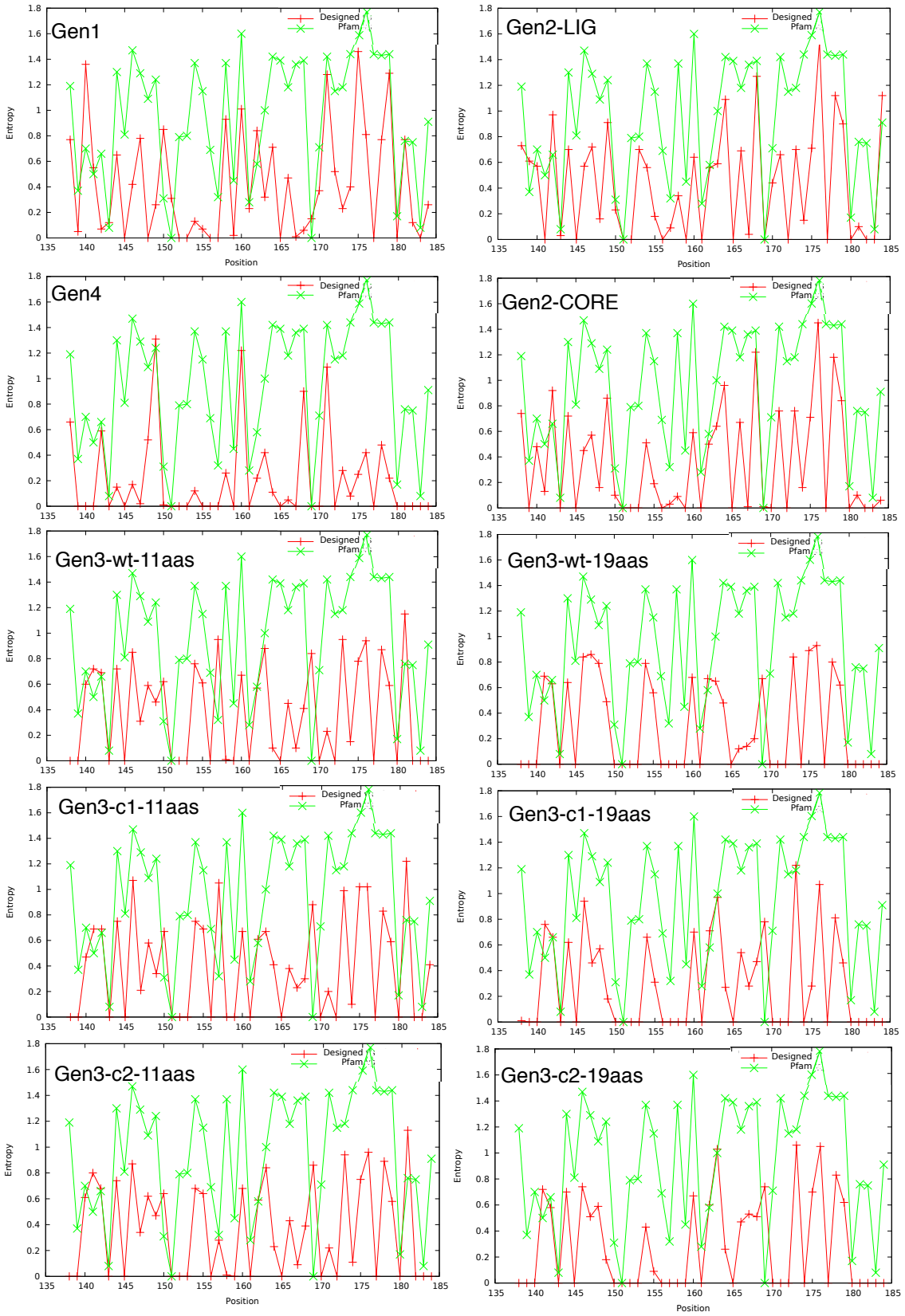


Figure 6.7 – Entropie moyenne par position pour chaque génération de séquences. Entropie moyenne par position pour les séquences de l'alignement PFAM.

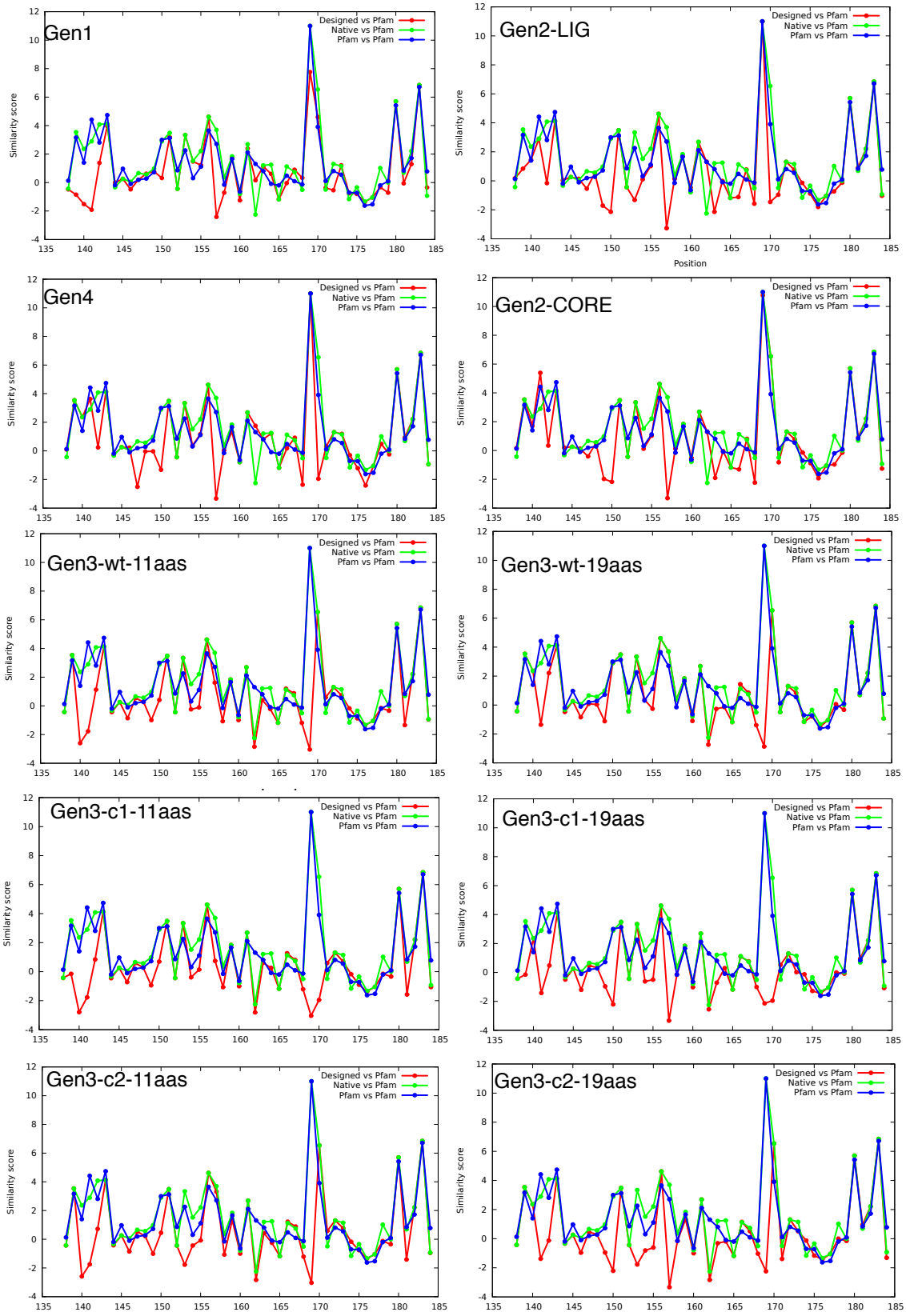


Figure 6.8 – Score de similarité (blosum) par position contre l'alignement PFAM et contre la séquence sauvage pour chaque génération de séquences. Score de similarité par position contre l'alignement PFAM pour la séquence sauvage.

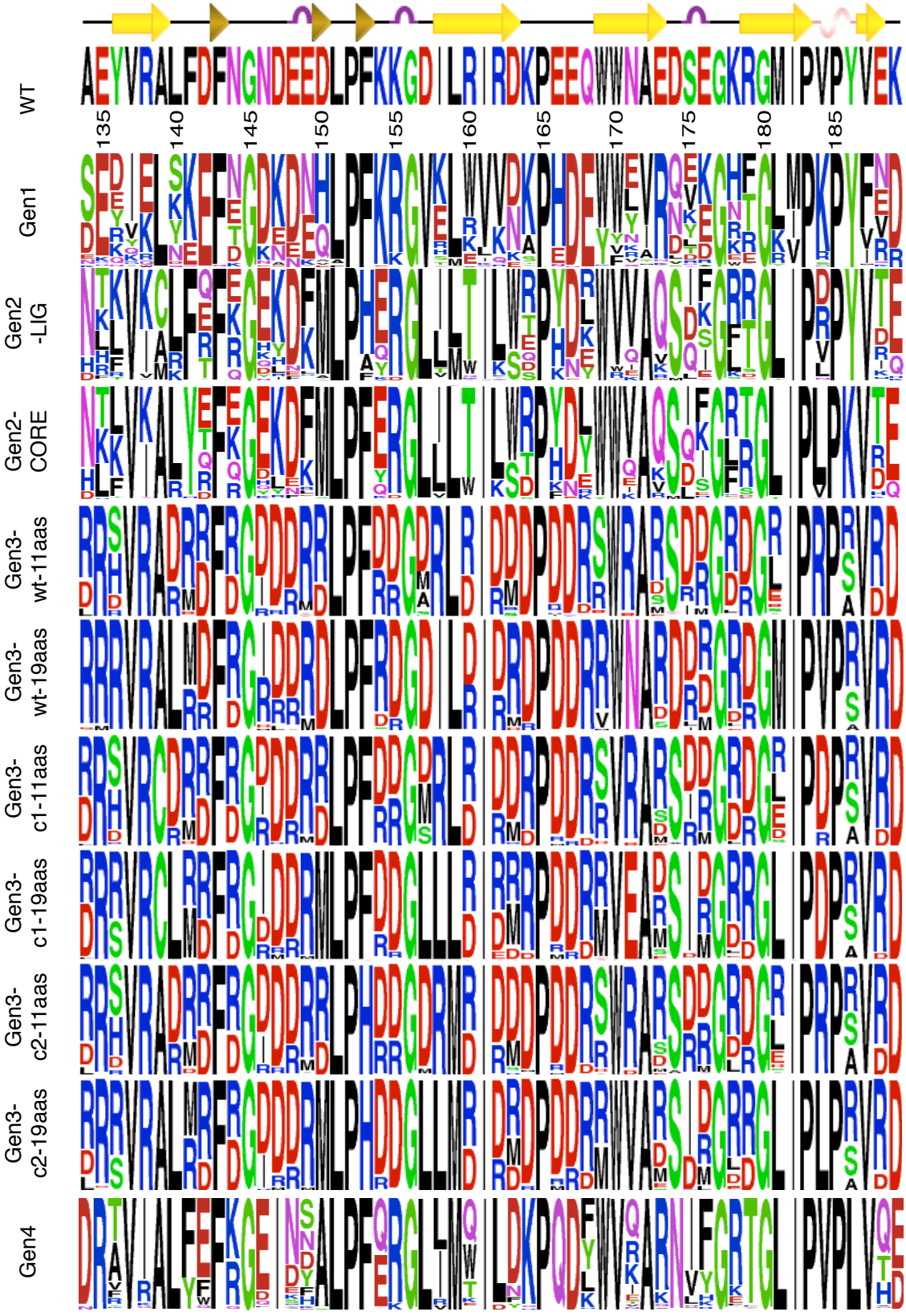


Figure 6.9 – Représentation logo (*weblogo*) de chaque génération de séquences. Les acides aminés sont représentés par leur code à une lettre. Et la taille des lettres est proportionnelle à la fréquence d'apparition de l'acide aminé correspondant dans l'alignement de séquences considéré à la position courante.

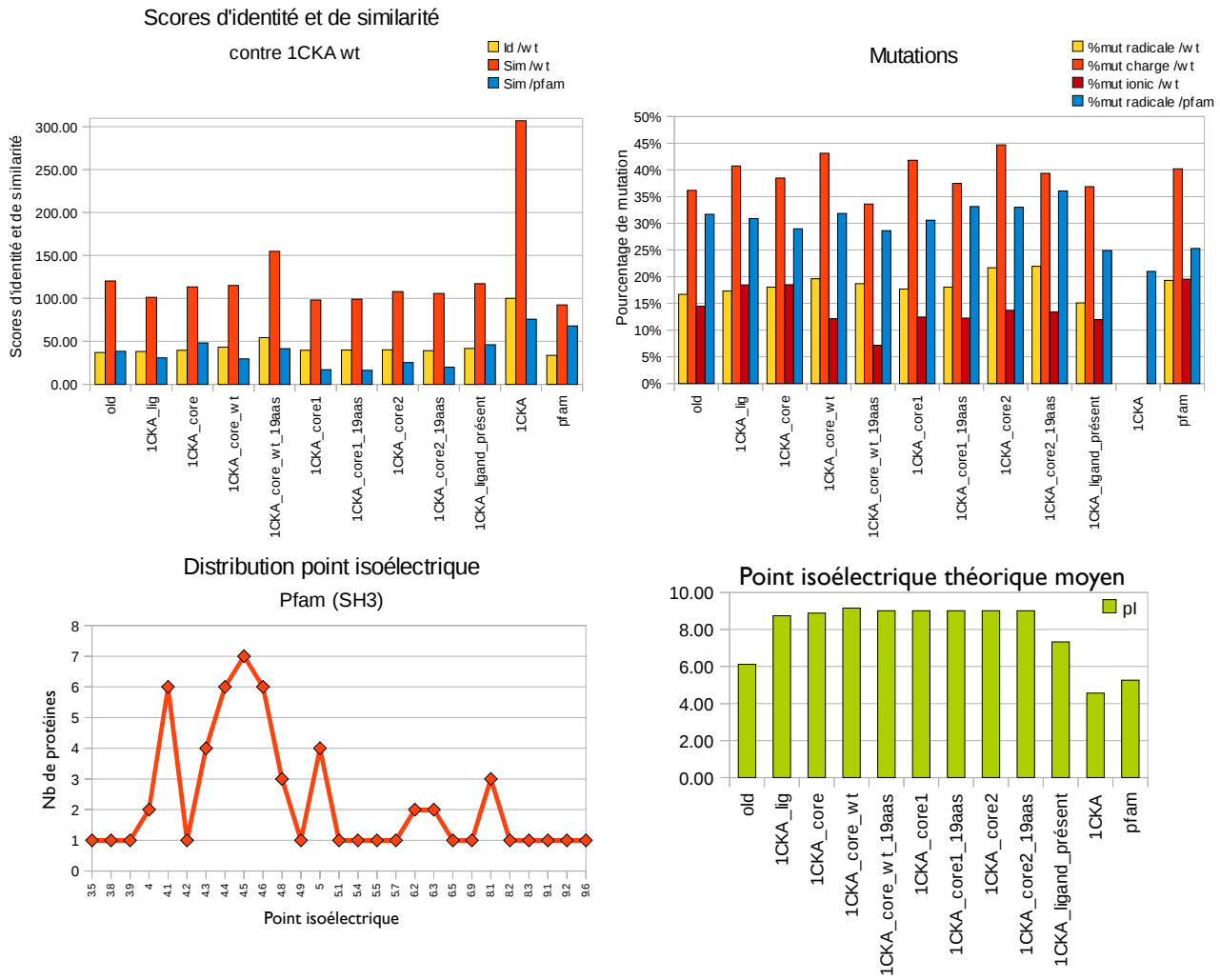


Figure 6.10 – Analyse des ensembles de séquences théoriques pour les différentes générations de séquences théoriques, ainsi que pour la séquence native 1CKA et la famille PFAM des domaines SH3. Distribution des points isoélectriques théoriques de l'alignement PFAM des domaines SH3

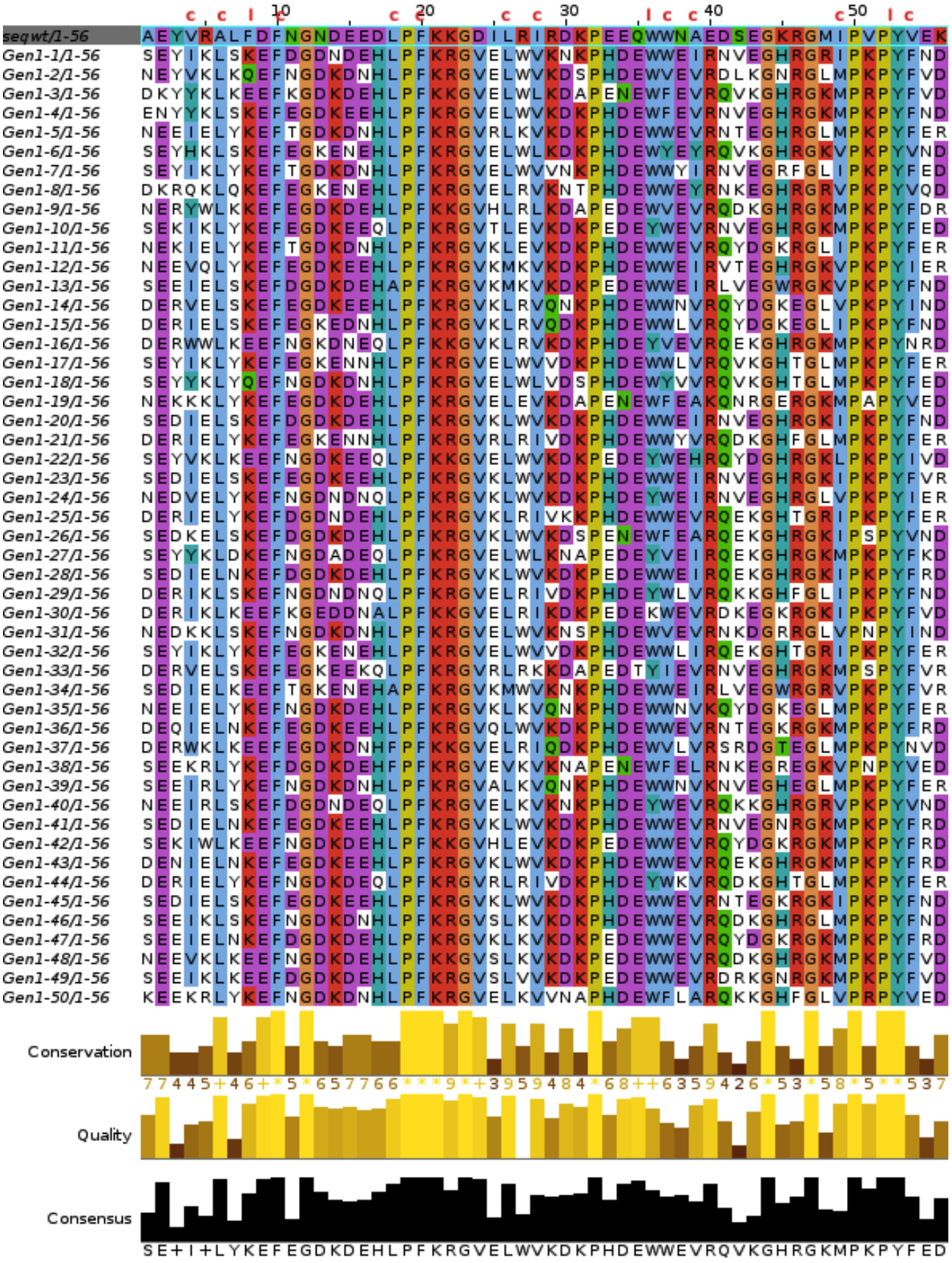


Figure 6.11 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen1-*old*.

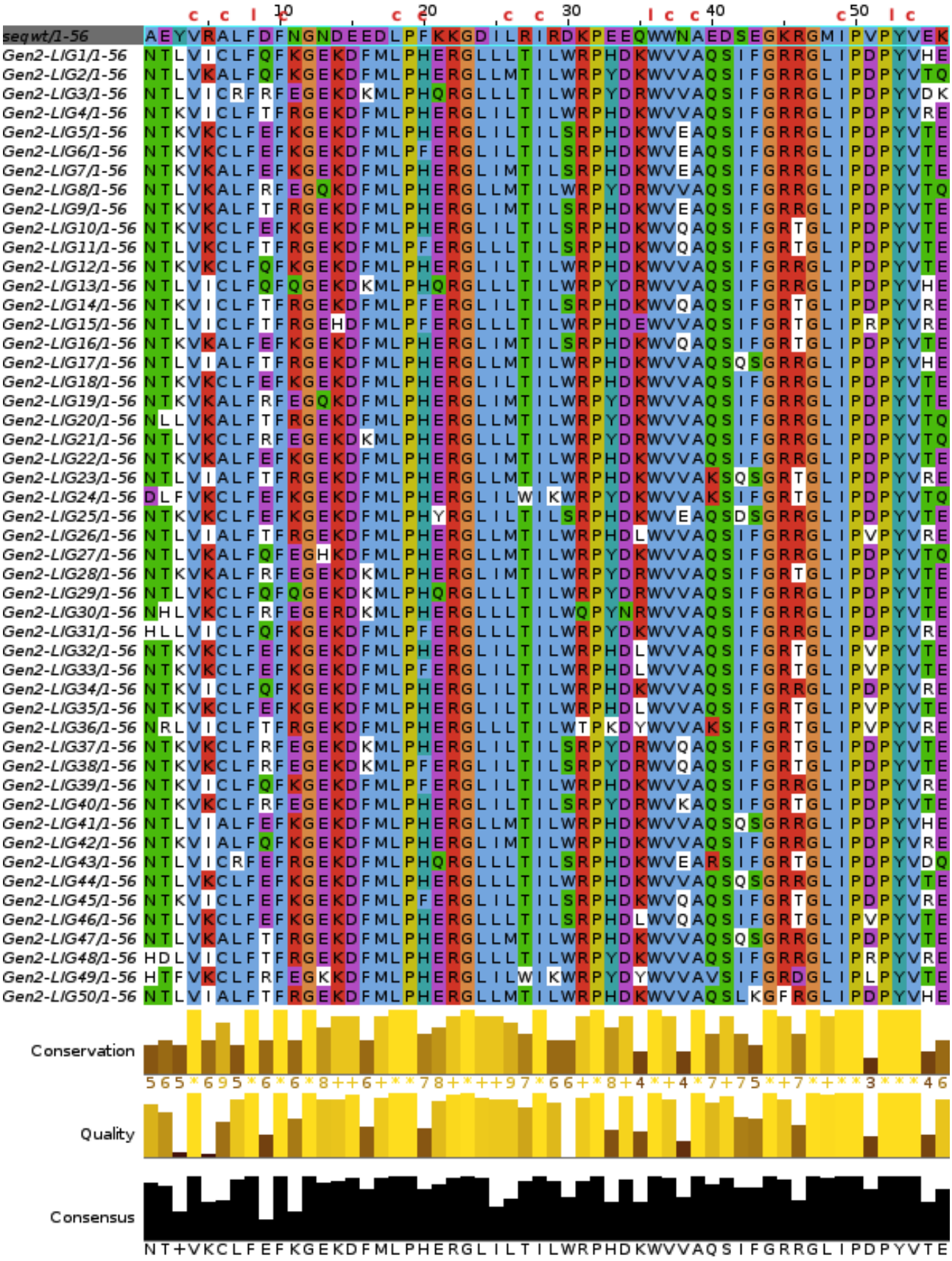


Figure 6.12 – Alignement des séquences théoriques (50 meilleures énergies Proteus)de la génération Gen2-LIG.

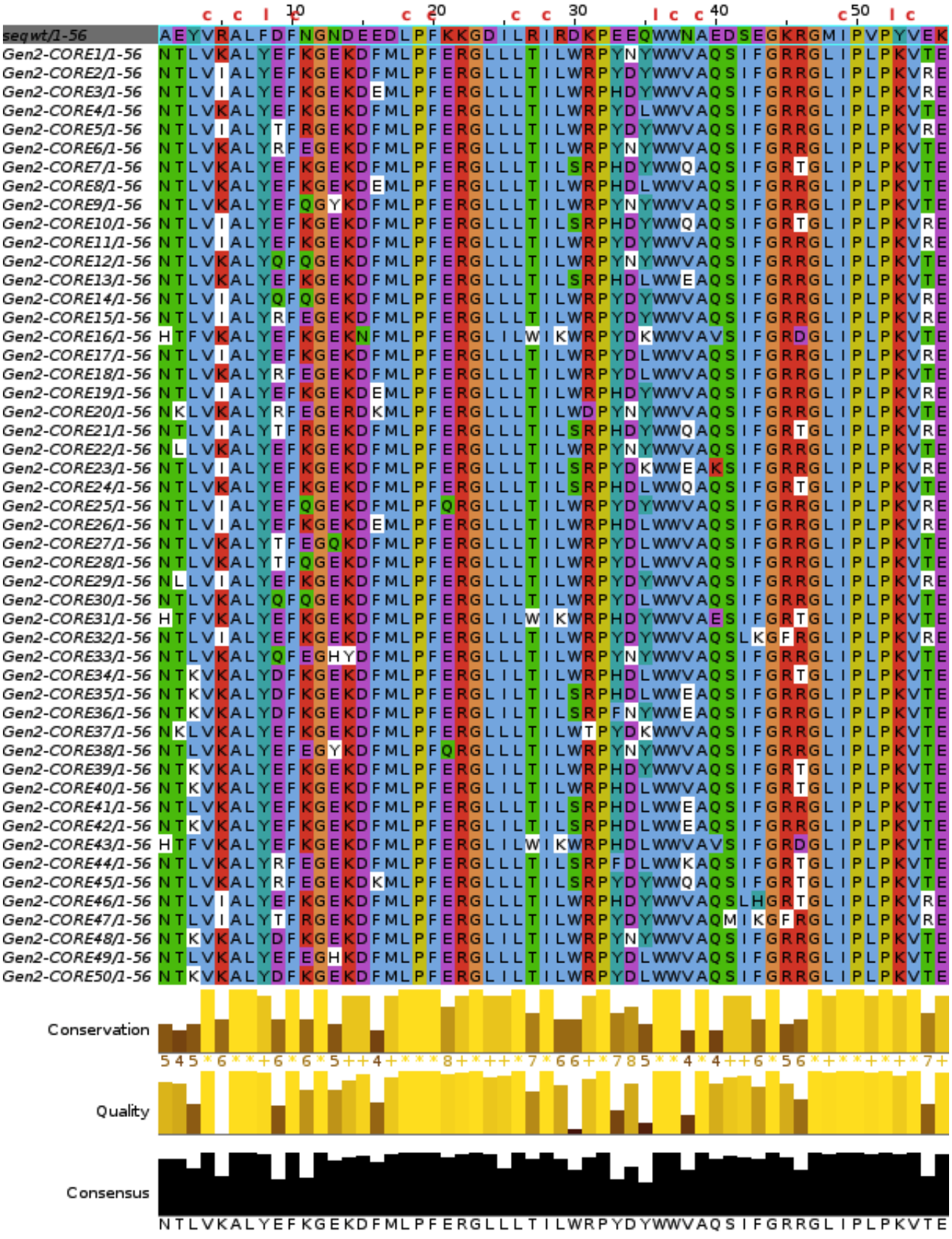


Figure 6.13 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen2-CORE.

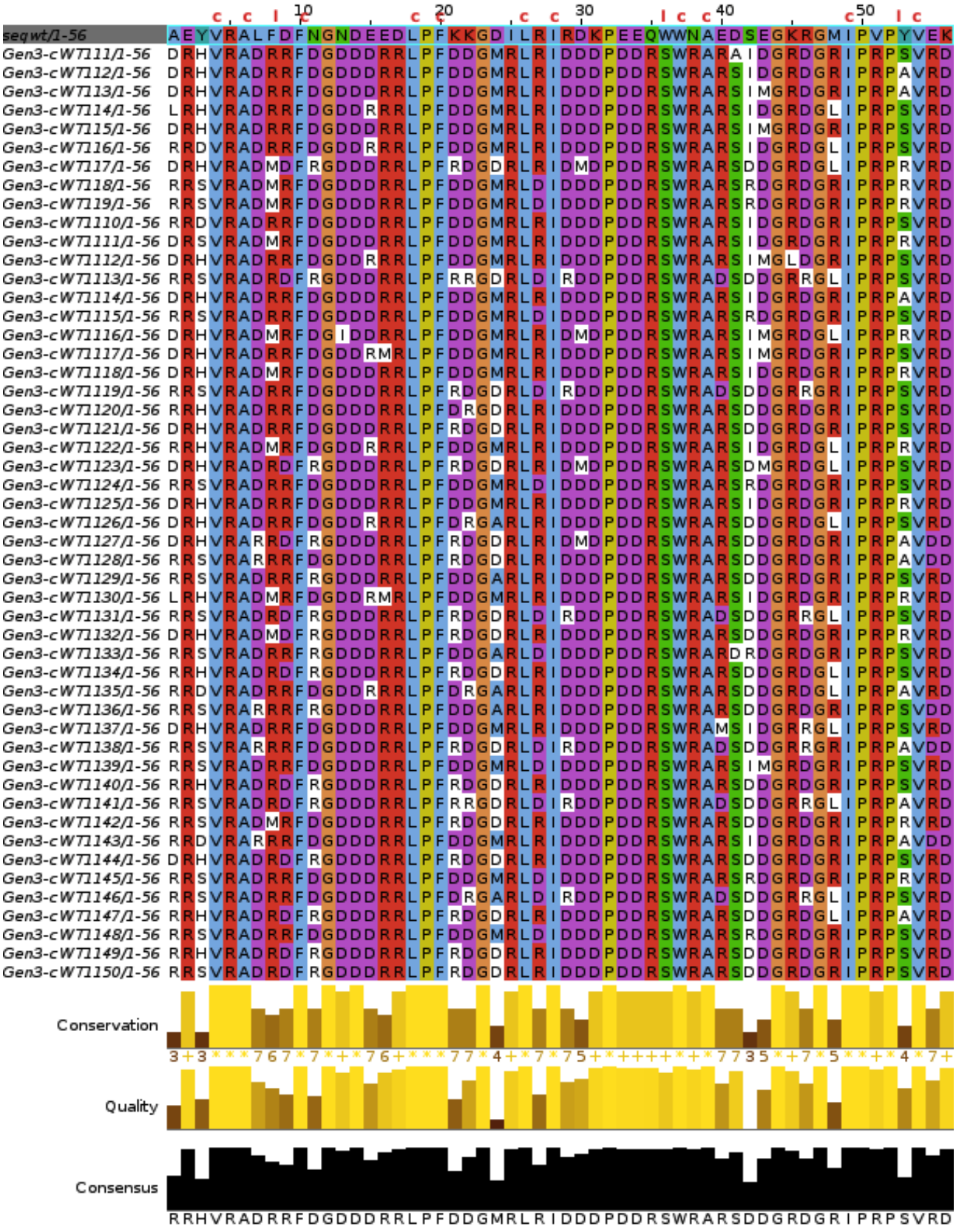


Figure 6.14 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-coreWT-11aas.

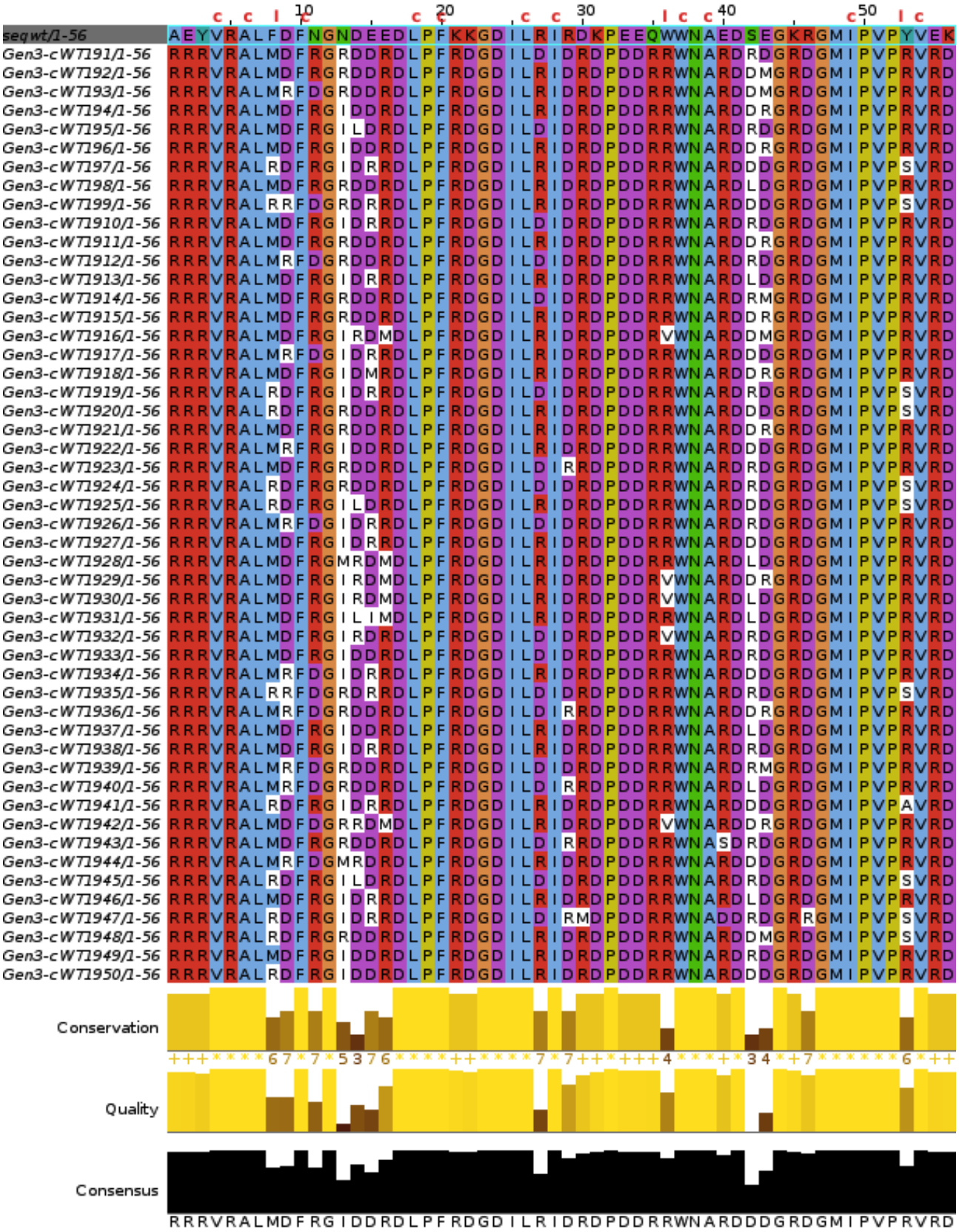


Figure 6.15 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-coreWT-19aas.

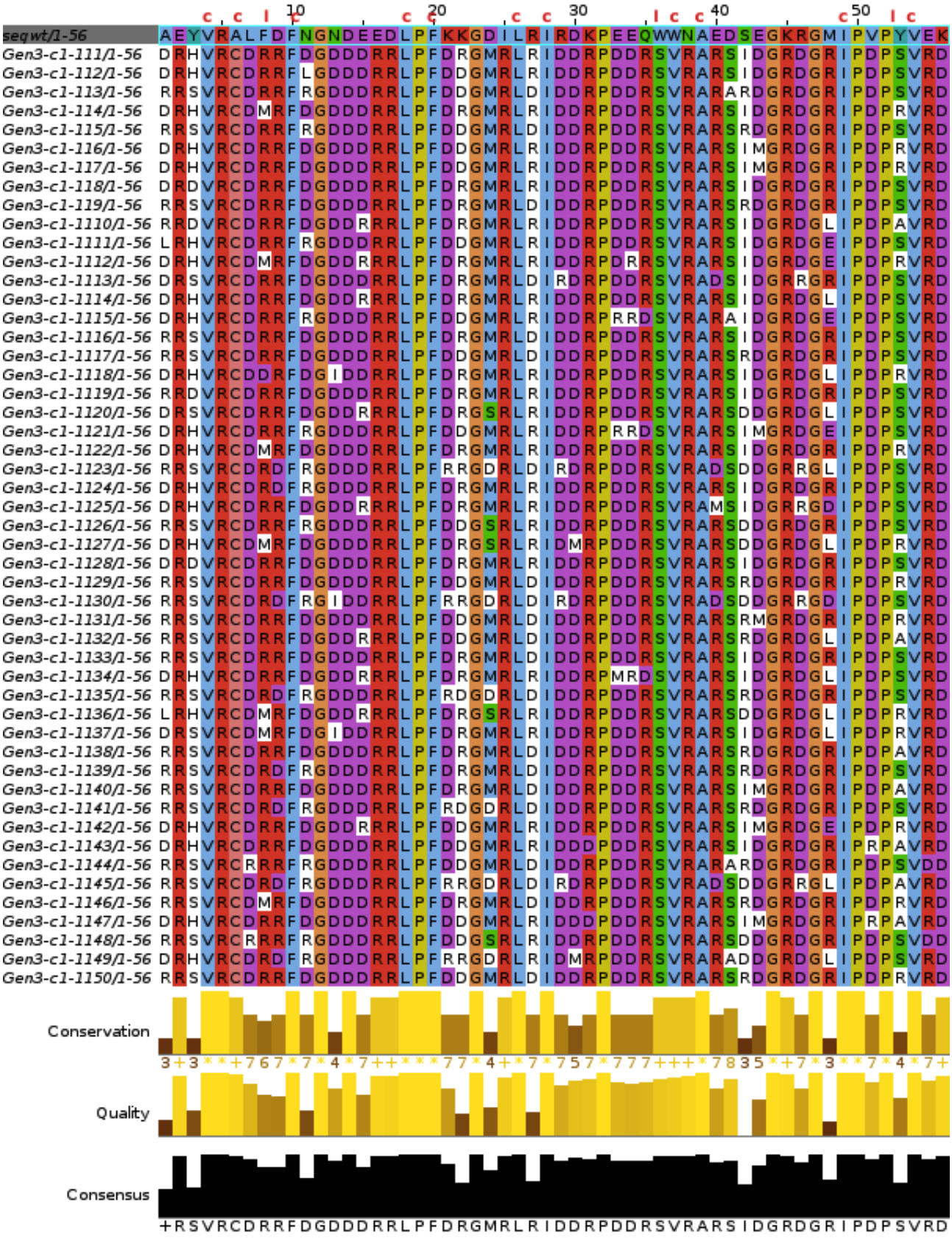


Figure 6.16 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-core1-11aas.

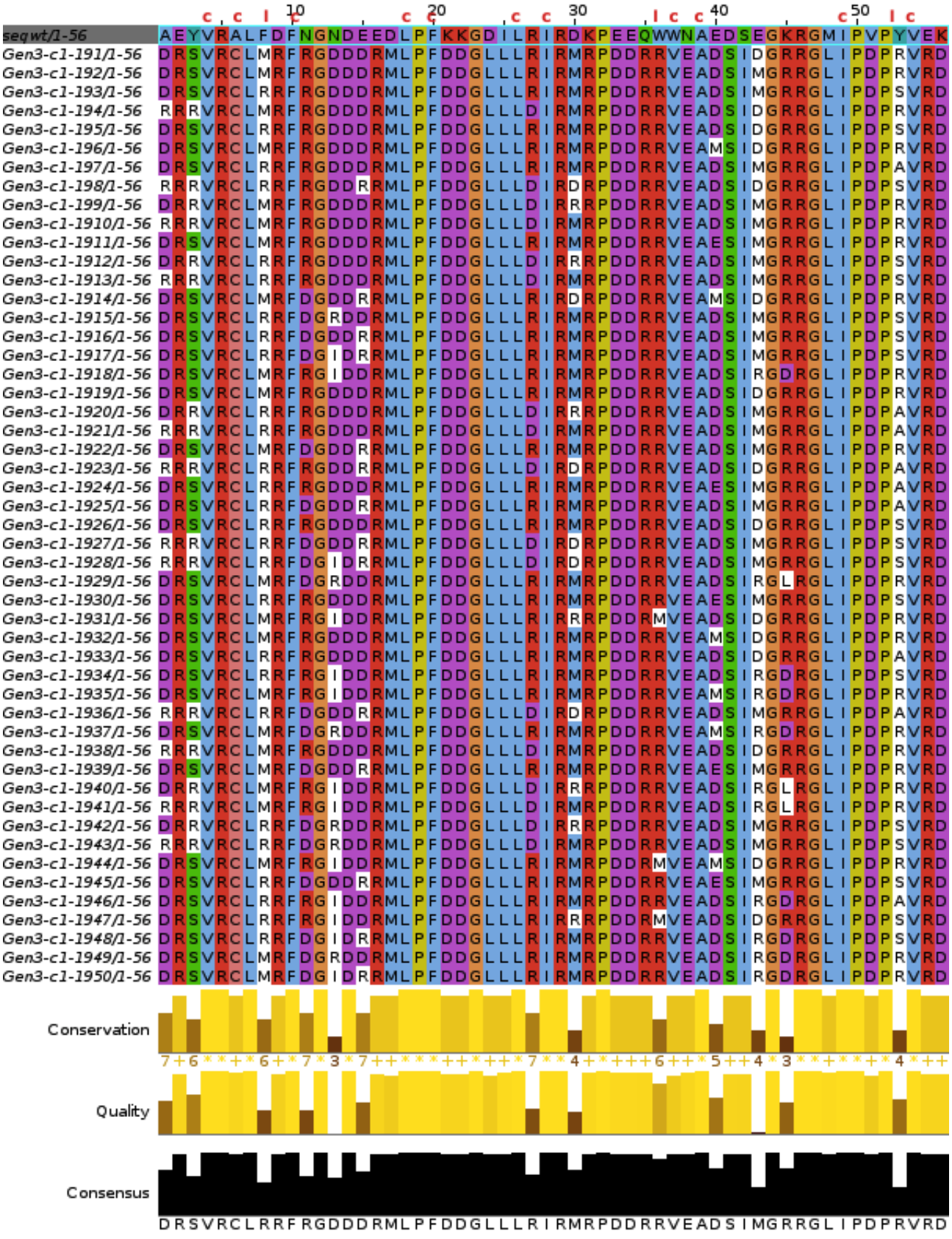


Figure 6.17 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-core1-19aas.

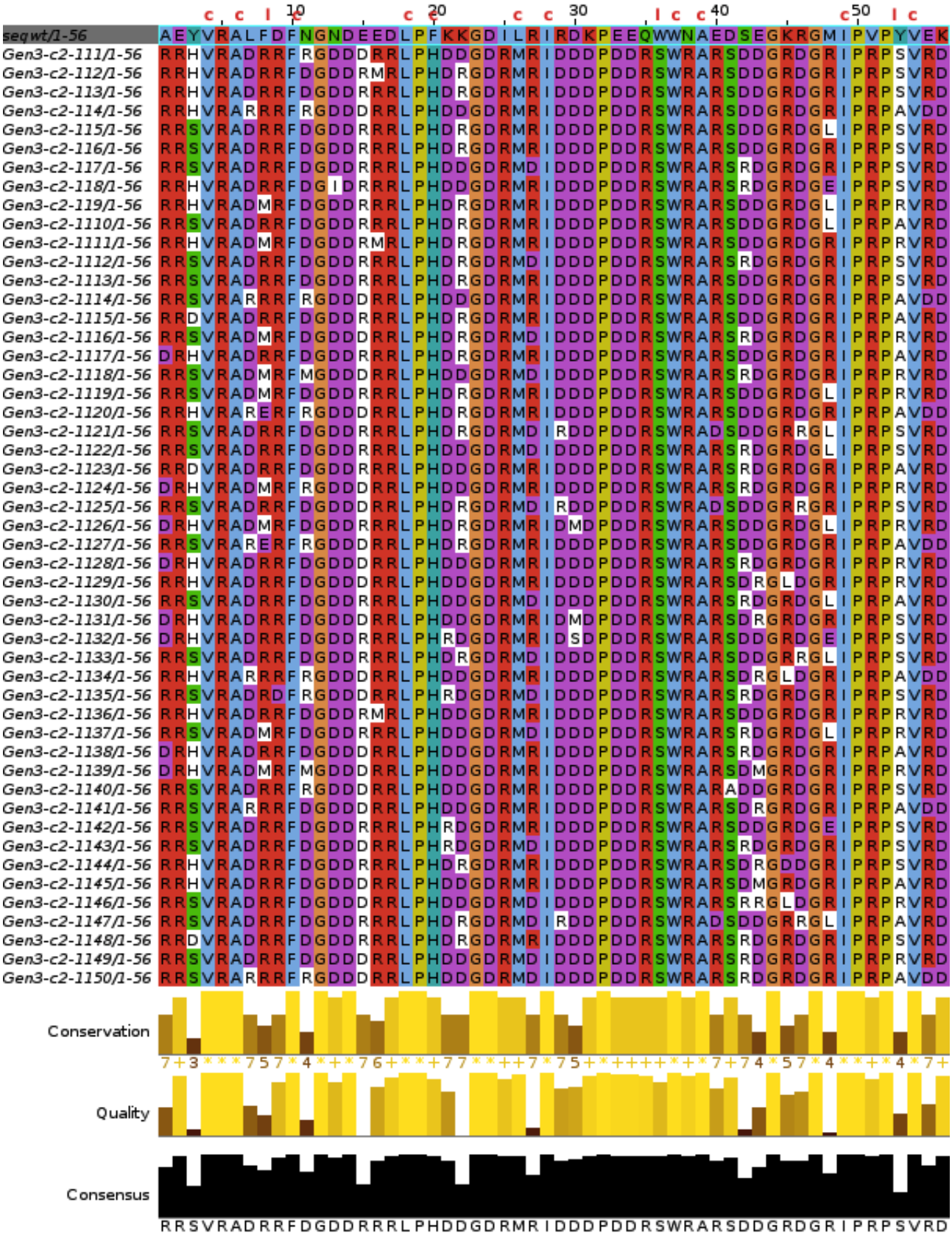


Figure 6.18 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-core2-11aas.

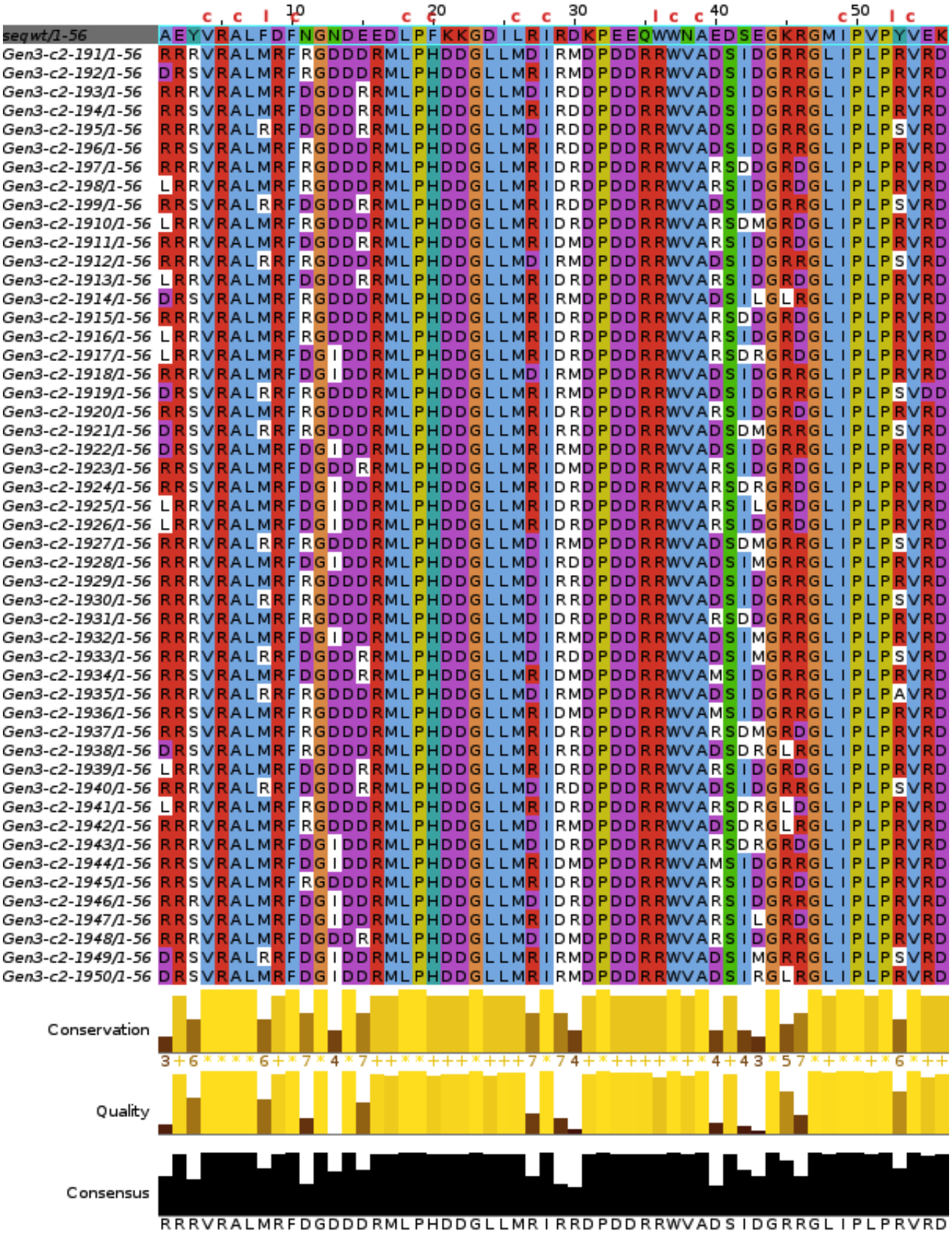


Figure 6.19 – Alignement des séquences théoriques (50 meilleures énergies Proteus) de la génération Gen3-core2-19aas.

6.4 Discussion et conclusion

6.4.1 Comparaison avec d'autres générations de séquences théoriques

Dans le cas de notre étude sur l'impact du coefficient surfacique uniforme sur la surface, aucun amélioration spectaculaire n'a été observée. Notre modèle initial "$\epsilon = 16$ et σ = PHIA" semble finalement être le meilleur compromis et donne des résultats assez satisfaisants.

Néanmoins, l'efficacité des modèles et des paramètres dépend surtout de l'objectif. Ce ne sera pas les mêmes problématiques, ni les mêmes attentes dans le cas d'une ingénierie de cœur hydrophobe, d'une interface protéine-protéine, ou d'une interface protéine-ligand. De manière générale, il demeure une difficulté à traiter de la même façon le cœur hydrophobe et la surface. Une solution serait de ne pas utiliser les mêmes jeux d'énergies de référence.

Par exemple, la méthode qui donne les meilleures prédictions pour les changements de stabilité utilise $\epsilon = 4$ ou 8 avec les coefficients surfaciques uniforme de $-0{,}04$ ou $-0{,}03$ [Lopes *et al.* 2007]. Avec ce modèle, les changements de stabilité ont été prédit à 1 kcal/mol près, avec une forte corrélation entre les expériences et les calculs. Dans cet étude, les coefficients surfaciques PHIA donnaient des erreurs nettement plus grandes, et surtout sans aucune corrélation entre expérience et calcul. La composition de la surface en acides aminés revenait à être tirée au hasard. Donc si on ignore l'électrostatique dans un cas comme celui-ci, où il s'agit de mutations d'acides aminés chargés, les erreurs seront inacceptables.

6.4.2 Énergies de références

Nous avons comparé les compositions globales en acides aminés des alignements Pfam [Punta *et al.* 2012] *seed* et *full* des familles SH3 (PF00018), SH3 (PF00017) et PDZ

(PF00595). Le tableau 6.8 contient des précisions sur les alignements Pfam. L'alignement *full* est généré par HMM à partir de l'alignement *seed*.

Family	Pfam	Nb seq.
SH3	*seed*	61
	full	8993
SH2	*seed*	58
	full	4403
PDZ	*seed*	58
	full	12568

Table 6.8 – Récapitulatif des alignements Pfam pour les familles de protéines SH3, SH2 et PDZ, avec le nom des alignements et le nombre de séquences qu'ils contiennent.

Nous rappelons que les énergies de références utilisées jusqu'à présent sont optimisées en prenant pour cible les abondances des acides aminés dans l'ensemble des alignements *seed* des familles SH3, SH2 et PDZ.

Dans la figure 6.20 en haut à gauche, on peut voir que pour la plupart des familles de protéines, les alignements Pfam *seed* et *full* ont une composition en acides aminés proche. En revanche, il y des disparités notables entre les trois familles de protéines. On observe par exemple que l'aspartate (D) et le glutamate (E) sont environ 2,5% plus abondants dans la famille SH3 que dans les autres familles de protéines.

Pour générer des séquences théoriques plus proches des protéines natives, il faudra peut-être choisir des énergies de référence ciblant les abondances en acides aminés de la famille concernée, ici les domaines SH3. En effet, comme le montre la figure 6.20 en bas, pour les générations Gen2 et Gen4, on a créé par rapport à la famille SH3, un déficit en acides aminés chargés négativement (Aspartate D et Glutamate E) et un excès en acide aminé chargé positivement (Arginine R) dans les séquences théoriques que nous générons. Alors que ces tendances sont inversées pour la Gen1. Ces déséquilibres d'abondance ont un impact direct sur les points isoélectriques des séquences prédites, comme on a pu le voir sur le tableau 6.10.

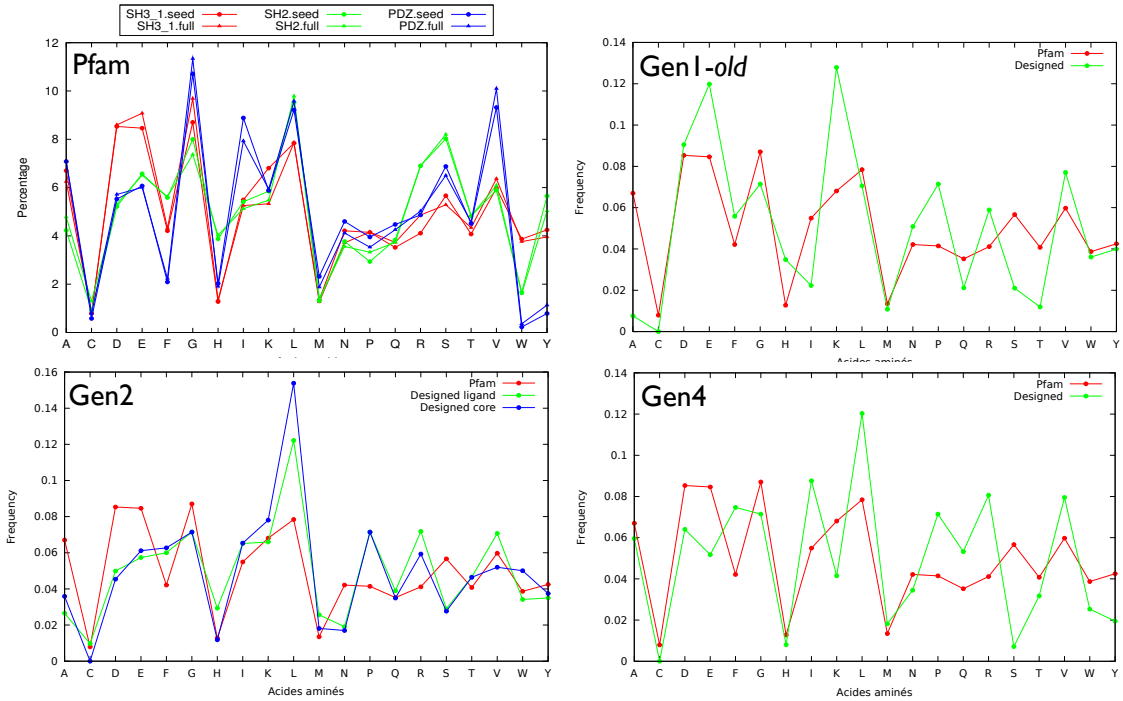


Figure 6.20 – Abondance des acides aminés.

Nous pouvons également se poser la question dans les cas des prédictions des résidus du cœur hydrophobe et ceux de la surface. Faut-il cibler encore plus rigoureusement les abondances en acides aminés particulières au cœur hydrophobe et à la surface ? Peut-être est-ce pousser la spécialisation du modèle un peu trop loin.

6.4.3 Conclusion

Nous pouvons conclure tout de même que nos séquences prédites sont de très bonne qualité par rapport aux séquences natives PFAM. Dans la figure 6.21, on peut constater que le cœur hydrophobe est très bien prédit, peu importe la génération. Les séquences prédites des générations Gen1 et Gen2-LIG et Gen4 ont des similarités contre PFAM assez bonnes.

Nous rappelons au lecteur que la zone d'ombre (*twilight zone*) représente un ensemble de protéines homologues par la structure et non par la séquence. Ainsi, vouloir être très

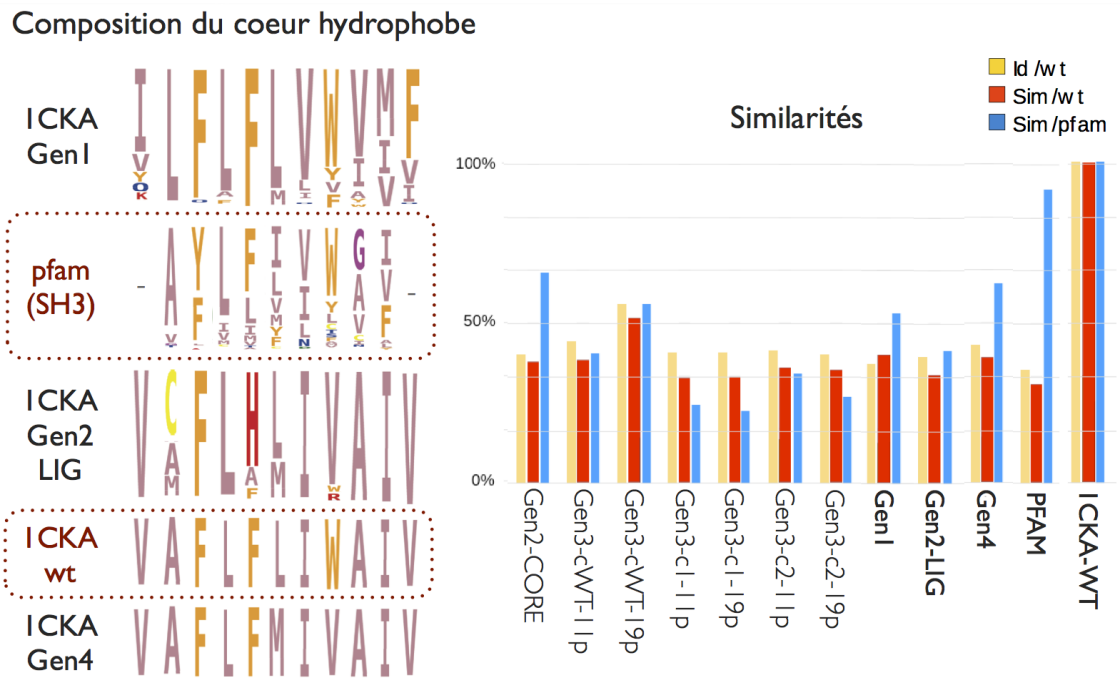


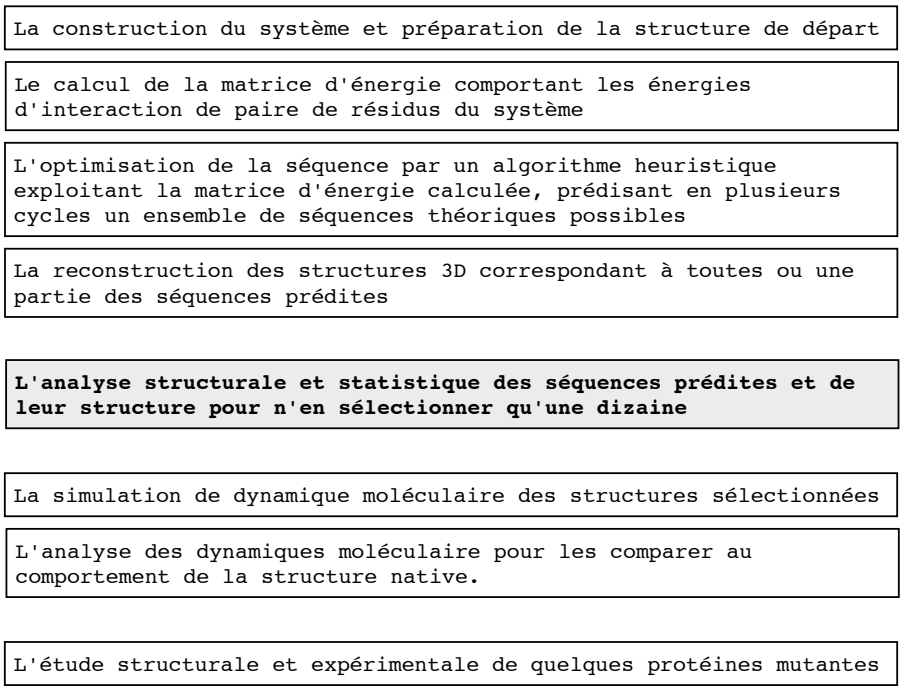
Figure 6.21 – Comparaison des cœurs hydrophobes prédits (format logo) et scores de similarité contre les séquences sauvages et PFAM.

proche des compositions en acides aminés d'une famille de protéines ou même d'une partie d'une protéine (cœur ou surface) pourrait écarter dans nos prédictions, des séquences plus éloignées mais dont la structure serait inchangée.

Néanmoins, prédire des séquences en respectant les caractéristiques diverses de la surface ou du cœur hydrophobe demeure un choix prudent mais sûr. Avant de vérifier la structure expérimentale des protéines mutantes, il convient cependant de sélectionner quelques séquences de bonne qualité. C'est que nous allons détaillé dans le chapitre suivant.

Chapitre 7

Mise en place de descripteurs pour l'analyse qualitative des séquences théoriques



Dans ce chapitre, nous décrirons donc le protocole mis en œuvre pendant cette thèse pour la caractérisation et la sélection des séquences les plus pertinentes et prometteuses par le biais de descripteurs divers (en gras ci-dessus).

Nous appliquerons principalement ces filtres sur les séquences théoriques des générations Gen2 (cas des positions fonctionnelles fixes et des positions du cœur hydrophobe fixes) et Gen4, afin de sélectionner des candidats. Pour pouvoir comparer l'impact de ces filtres, nous les avons appliqués également à la génération Gen1-*old*, ainsi que sur les six générations Gen3.

Les simulations de dynamique moléculaire seront décrites et analysées dans le chapitre suivant.

7.1 Choix de séquences théoriques candidates à tester en dynamique moléculaire

Parmi plusieurs milliers de séquences théoriques que nous générons, comment choisir de "bons" candidats pour les tester expérimentalement ensuite ? Comment définir une "bonne" séquence ? Sur quels critères devons-nous nous baser pour faire ce choix ? Nous avons mis en place un protocole de tri utilisant des descripteurs suffisamment riches en information pour nous permettre de caractériser finement nos séquences théoriques, que nous qualifierons également par la suite de mutantes.

7.1.1 Protocole

Un premier filtre est appliqué pour éliminer les séquences de trop haute énergie PROTEUS. La séquence de plus basse énergie est prise comme référence et toutes les conformations dont l'énergie lui est supérieure de plus de 5 kcal/mol sont éliminées. Plus arbitrairement mais tout de même en accord avec ce seuil, nous prenons les 5000 séquences de meilleure énergie.

Ensuite, les conformations restantes sont comparées par similarité de volume structural du cœur hydrophobe avec celui de la protéine native. La déviation structurale autorisée entre les cœurs hydrophobes mutants et le cœur natif est définie par une valeur

seuil : $\pm 100\,\text{Å}^3$. Différentes valeurs furent testées, néanmoins un seuil un peu plus tolérant conduit rapidement à un nombre de séquences trop important. Nous réduisons d'un facteur 10 le nombre de séquences considérées après ce premier filtre.

Puis, nous comparons la distribution des scores de similarité Blosum entre nos séquences théoriques et l'alignement PFAM des domaines SH3.*seed* et la distribution des scores de similarité de l'alignement PFAM et lui-même. Nous pouvons ainsi décidé d'un seuil minimum subjectif et heuristique pour le score de similarité. Cela nous permet de réduire encore le nombre de séquences à considérer.

Un dernier filtre strict est alors appliqué : nous calculons le nombre de mutations radicales (que nous définirons un peu plus tard) sur nos séquences théoriques par rapport à l'alignement PFAM des domaines SH3. En le comparant également à la distribution du nombre de mutations radicales au sein de l'alignement PFAM, on peut déterminer un seuil maximum de mutations radicales acceptables.

Les deux dernières étapes de filtration sont beaucoup plus manuelles. Nous considérons tout d'abord le point isoélectrique (pI) théoriques de chaque séquence, afin d'éliminer les protéines qui pourraient nous poser des problèmes lors de la production dans *Escherichia coli in vivo*. En effet, les protéines dont le point isolélectrique s'approche de 7, auront un risque d'agrégation en milieu physiologique.

Enfin, nous étudions les fréquences d'apparition de certains motifs d'acides aminés sur quelques positions afin de garder une variabilité dans les séquences mutantes finales sélectionnées et/ou de se rapprocher de la séquence native. La sélection d'une dizaine de séquences restent complètement "manuelle" à ce niveau.

Voir le schéma 7.10 dans la partie "Résultats".

7.1.2 Descripteurs

Afin dans un premier temps d'évaluer la qualité des séquences prédites, et dans un deuxième temps de sélectionner les meilleures séquences, nous avons mis en place une

série de descripteurs. Ces descripteurs se basent sur des comparaisons entre nos séquences théoriques et soit la séquence native, soit des alignements multiples (MSA) de séquences expérimentales provenant de la base de donnée PFAM. Certains de ces descripteurs, les plus pertinents, serviront à terme de filtres pour sélectionner les séquences théoriques les plus prometteuses pour les tests expérimentaux. Dans le tableau 7.1, on peut voir les différents descripteurs considérés et leur type de comparaison.

Descripteurs	PFAM/PFAM	WT/PFAM	WT/théoriques	Théoriques/PFAM
Score d'identité *	S/P	S/P	S/P	S/P
Score de similarité *	S/P	S/P	S/P	S/P
Nb mutations radicales *	S/P	S/P	S/P	S/P
Charge globale *	S	S	S	S
Nb mutations de charge	S	S	S	S
Masse / Masse native	S	S	S	S
Volume /volume natif *	S	S	S	S
Énergie	S	S	S	S
Point isoélectrique *	S	S	S	S
Entropie	P	P	P	P
Composition en Aas	P	P	P	P

Table 7.1 – Descripteurs utilisés sur les séquences théoriques. S : comparaison par séquences, P : comparaison par positions. Les descripteurs avec une * son détaillés un peu plus loin.

Pour le calcul de charge, de masse, de volume et d'énergie, nous avons utilisé des paramètres propres à chaque type d'acide aminé. Nous les avons répertorié dans le tableau 7.2.

Les descripteurs utilisent pour comparer les séquences théoriques et les séquences expérimentales des domaines SH3 connus, l'alignement multiple PFAM représenté dans la figure 7.1.

7.1.2.1 Score d'identité

Le score d'identité est simplement le taux d'identité en terme d'acide aminé. On peut noter plus rigoureusement :

$$s(x,y) = \sum_{1 \leq i \leq n} s(x_i,y_i) \tag{7.1}$$

Acide aminé	Charge	Masse	Volume	Énergie de référence
ALA (A)	0	89.0940	34.092	-10.052
ARG (R)	1	174.2027	132.740	-22.278
ASN (N)	0	132.1190	75.372	-17.438
ASP (D)	1	133.1038	71.898	-20.713
CYS (C)	0	121.1540	56.225	0.000
GLN (Q)	0	146.1459	97.405	-18.439
GLU (E)	1	147.1307	93.931	-20.787
GLY (G)	0	75.0671	0.000	0.000
HIS (H)	1	155.1563	104.420	-18.269
ILE (I)	0	131.1746	102.647	-10.255
LEU (L)	0	131.1746	102.647	-12.305
LYS (K)	1	146.1893	108.755	-21.628
MET (M)	0	149.2078	105.121	-12.557
PHE (F)	0	165.1918	135.823	-18.246
PRO (P)	0	115.1319	66.099	0.000
SER (S)	0	105.0934	45.299	-12.812
THR (T)	0	119.1203	69.788	-11.840
TRP (W)	0	204.2284	168.557	-18.881
TYR (Y)	0	181.1912	147.005	-20.988
VAL (V)	0	117.1478	80.614	-9.743
Gap (-)	0	0.0000	0.000	0.000

Table 7.2 – Paramètres pour chaque type d'acide aminé utilisés dans les calculs de plusieurs descripteurs.

où x_i est l'acide aminé (ou gap) à la position i de la séquence concernée, et y_i l'acide aminé (ou gap) à la position i de la séquence comparée. On définit $s(x_i,y_i) = 1$ si les deux séquences ont le même acide aminé en position i. Sinon $s(x_i,y_i) = 0$.

7.1.2.2 Score de similarité

Le principe est d'associer un score de similarité à un alignement de plusieurs séquences, plus particulièrement de deux séquences. Pour tenir compte des ressemblances et différences de propriétés physico-chimiques entre acides aminés, ainsi que de leur abondance dans la nature, on utilise une matrice de comparaison. La valeur du coefficient $M(a,b)$ reflète la qualité de l'alignement entre les deux acides aminés a et b. On peut donc calculer un score global pour l'alignement d'une séquence polypeptidique entière de longueur L :

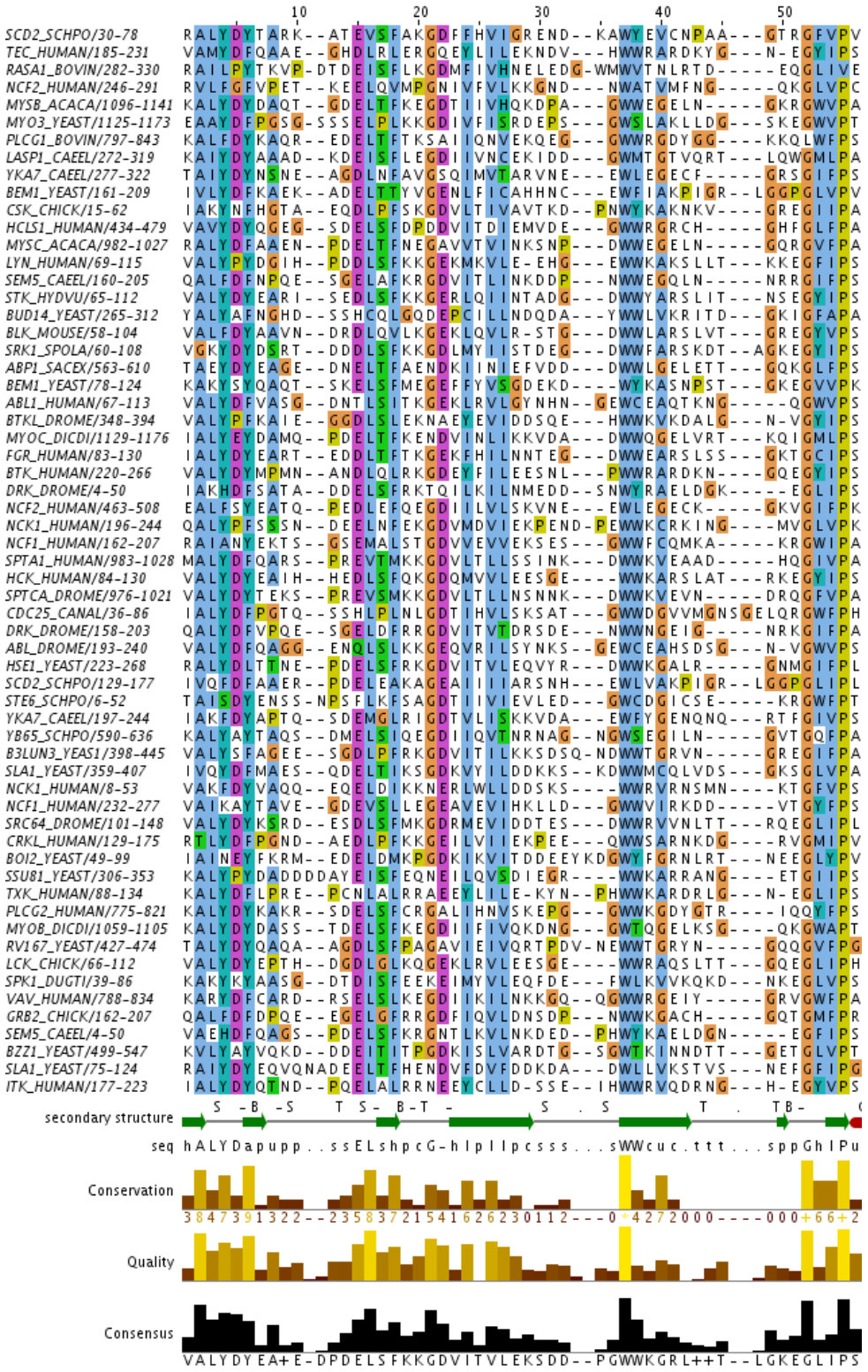


Figure 7.1 – Alignement multiple des séquences naturelles PFAM des domaines SH3.

$$score = \sum_{k=1}^{L} M(a_k,b_k) \tag{7.2}$$

Plusieurs matrices Blosum ont été calculées à partir de blocs d'alignement construits sur des critères d'identité, mais nous avons choisi la matrice Blosum62 (voir figure 7.2), construite sur des alignements de séquences à 62% identiques [Henikoff & Henikoff 1992 ; Sean 2004].

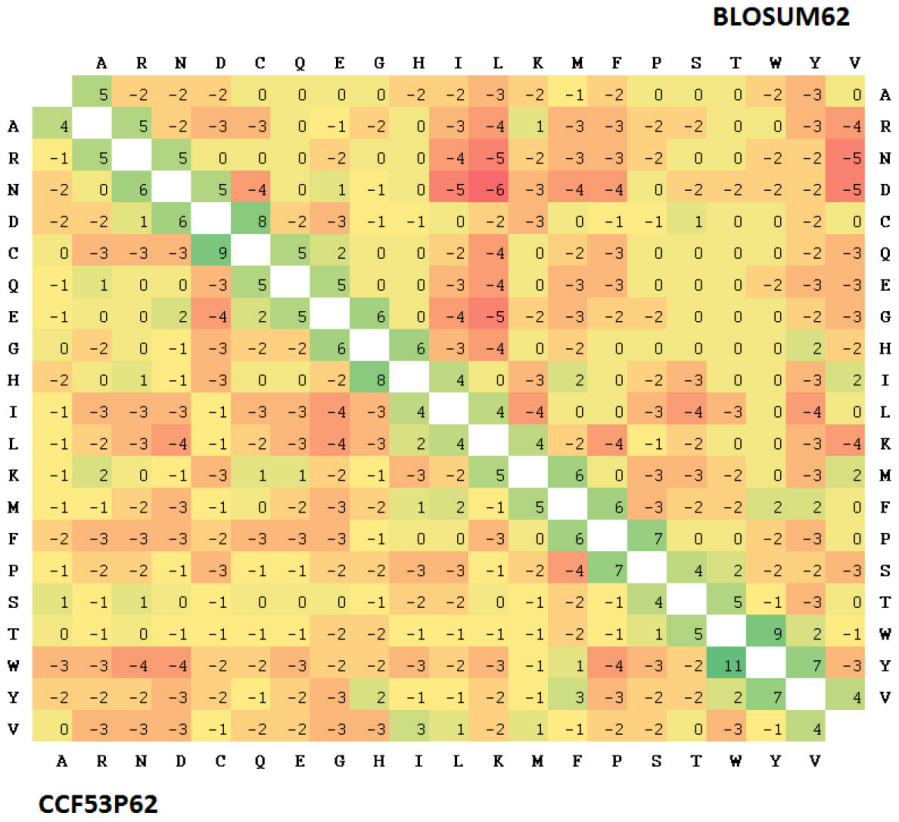


Figure 7.2 – Matrice Blosum62. On peut y voir en rouge les mutations les plus défavorables ou éloignées, et en vert les mutations favorables ou quasi silencieuses.

7.1.2.3 Mutations radicales

Nous définissons la notion de mutation radicale par les mutations entre deux acides aminés a et b dont le score dans la matrice Blosum62 est inférieur ou égal à -2. Ce qui

revient à caractériser les mutations qui changent radicalement une ou plusieurs propriétés physico-chimiques, de taille ou autres entre deux résidus, et ainsi à limiter leur nombre au sein de nos séquences mutantes par rapport à la séquence native.

7.1.2.4 Charge

La charge totale des séquences prédites est calculé par la formule :

$$Charge = \sum_{1 \leq i \leq n} c_i \tag{7.3}$$

avec c_i : charge d'acide aminé de la position i.

7.1.2.5 Volume structural du cœur hydrophobe

Dans le cadre de l'analyse structurale des protéines, la tessellation de Voronoï a été utilisée pour la première fois par Richards [1974] pour évaluer, dans une protéine globulaire, les volumes des atomes, définis par les volumes de leurs polyèdres de Voronoï. Étant donné un ensemble de points centroïdes, la tessellation "classique" de Voronoï divise l'espace en régions, appelées cellules de Voronoï, centrées sur ces points [Okabe *et al.* 2000]. Beaucoup d'améliorations et de méthodes dérivées ont été réalisées et analysées formellement (par exemple, par Edelsbrunner *et al.* [1996]). Toutes ces constructions ont permis de montrer que la tessellation de Voronoï est un bon modèle mathématique de la structure des protéines. Elle permet de montrer que les protéines sont des objets compacts et d'analyser les cavités dans les structures [Bakowies & Gunsteren 2002 ; Liang & Dill 2001 ; Peters *et al.* 1996]. Aussi, nous avons utilisé cette notion de volume structural pour quantifier les changements structuraux au sein des cœurs hydrophobes de nos protéines.

7.1.2.6 Point isoélectrique

Le point isoélectrique (pI) d'une protéine est défini comme étant le pH pour lequel la protéine est électriquement neutre ; on parle de sa forme zwitterionique. La valeur réelle du pI dépend de la composition en acides aminés, plus précisément des pKa des résidus chargés, de la structure tridimensionnelle et du tampon utilisé. Néanmoins, le calcul du pI réalisé ici est théorique et ne dépend que de la composition en acides aminés. Si pH $<$ pI, la charge globale est positive, car la molécule a tendance à conserver ses protons ou à en capter du milieu acide. Si pH $>$ pI, la charge globale est négative, car la molécule a tendance à céder ses protons au milieu basique (voir schéma 7.3).

$$pI = \frac{\sum_{i=1}^{n} pKa_i}{n} \tag{7.4}$$

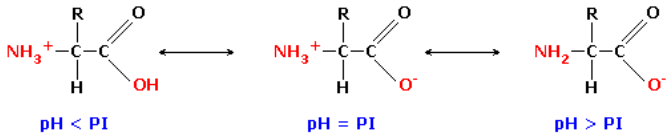


Figure 7.3 – Un acide aminé dans sa forme zwitterionique (pH = pI) et ionisée (pH <> pI).

7.2 Résultats pour les différentes générations de séquences

Les résultats de chaque filtre sont représentés dans les figures 7.4 à 7.8 pour chaque génération de séquences.

On peut observer dans la figure 7.5 que les générations Gen2-CORE, Gen3-WT-19aas et Gen4 ont les meilleurs scores de similarité et qu'ils recouvrent le bas de la distribution des scores de la famille PFAM. Ces séquences sont donc plus proches des séquences natu-

relles PFAM les plus variables. Le nombre de mutations radicales est directement lié au score de similarité, comme on peut le voir dans la figure 7.6.

Dans l'alignement de séquences (figure 7.8), on peut voir qu'il y a des zones mieux prédites par rapport à la séquence sauvage. Les générations Gen3 et Gen4 prédisent bien mieux les acides aminé chargés négativement (D et E), par exemple dans la zone 163-167, par rapport aux générations Gen1 et Gen2. Cela est étroitement lié avec l'abondance des acides aminés et leurs énergies de référence.

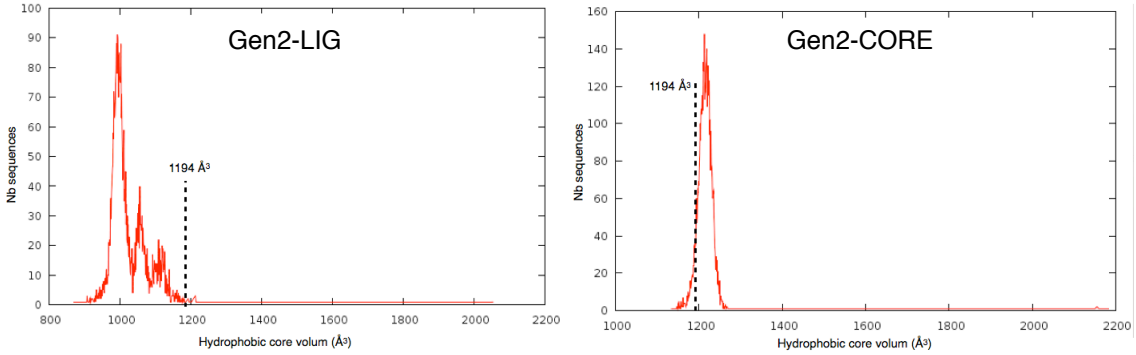


Figure 7.4 – Résultats des descripteurs : volume du cœur hydrophobe pour les séquences prédites Gen2 avec les positions fonctionnelles et du cœur hydrophobe fixes.

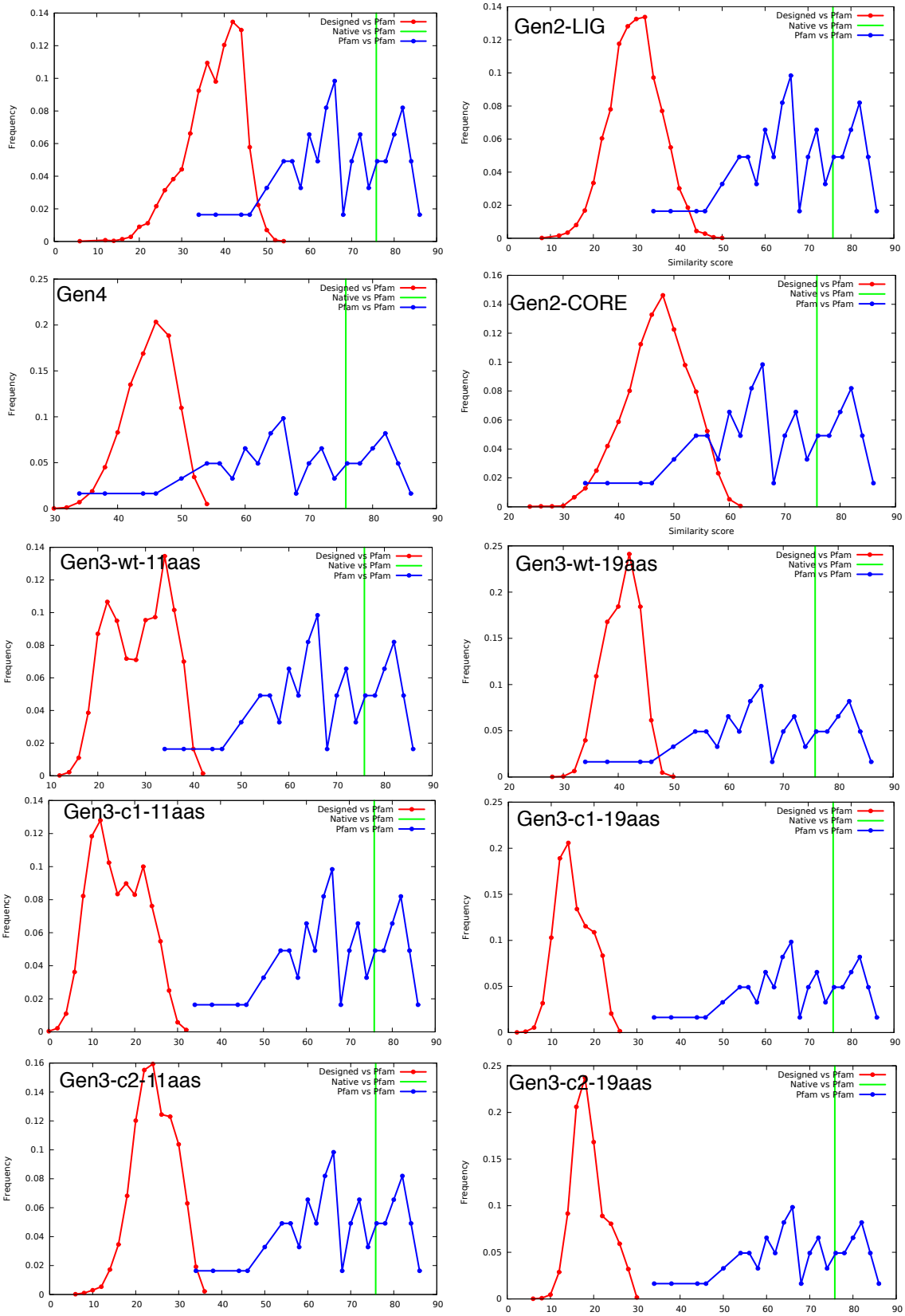


Figure 7.5 – Résultats des descripteurs : score de similarité (Blosum) contre l'alignement PFAM pour toutes les générations de séquences (de Gen1 à Gen3.

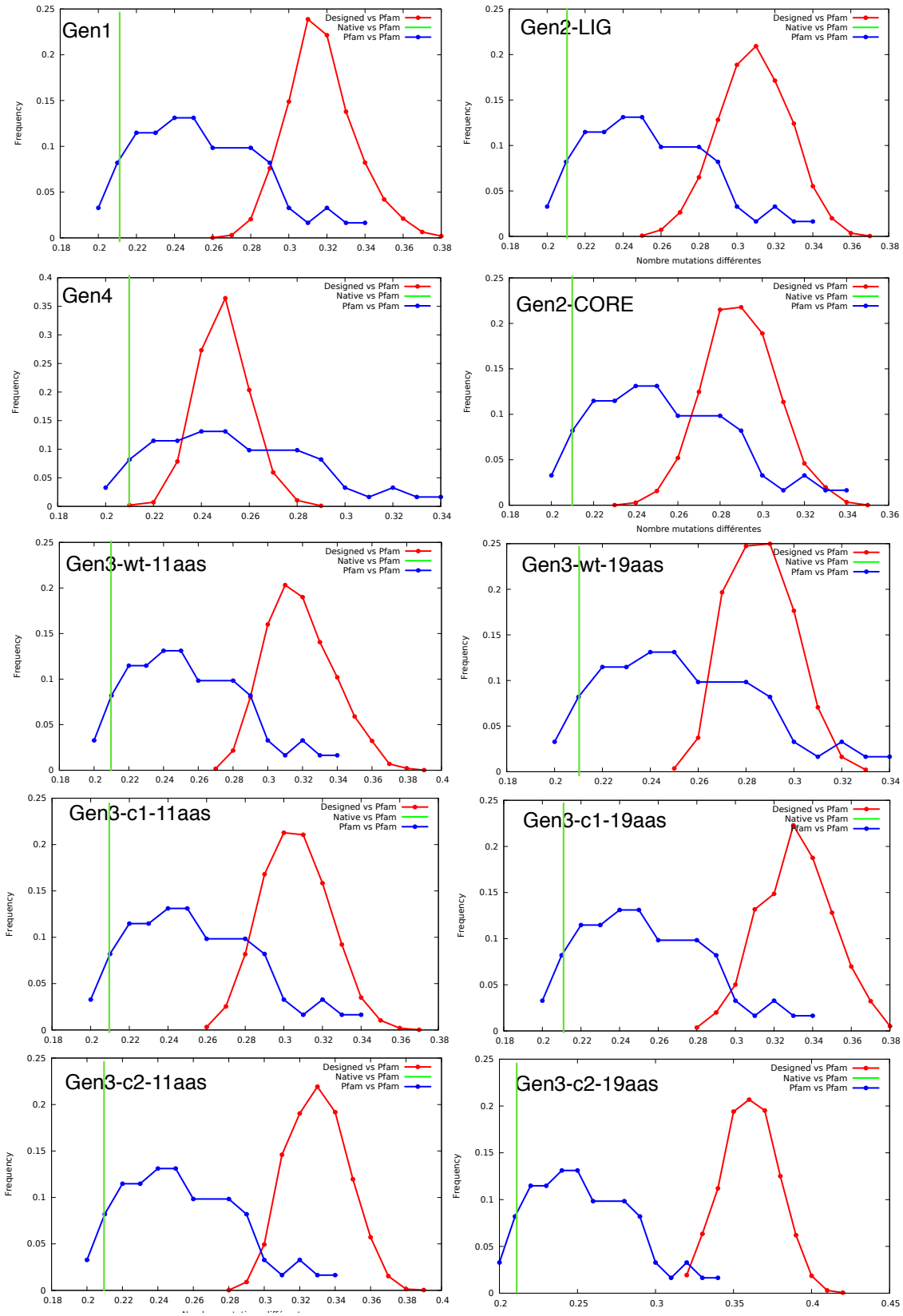


Figure 7.6 – Résultats des descripteurs : pourcentage de mutations radicales par rapport à l'alignement PFAM dans les séquences prédites (de Gen1 à Gen3).

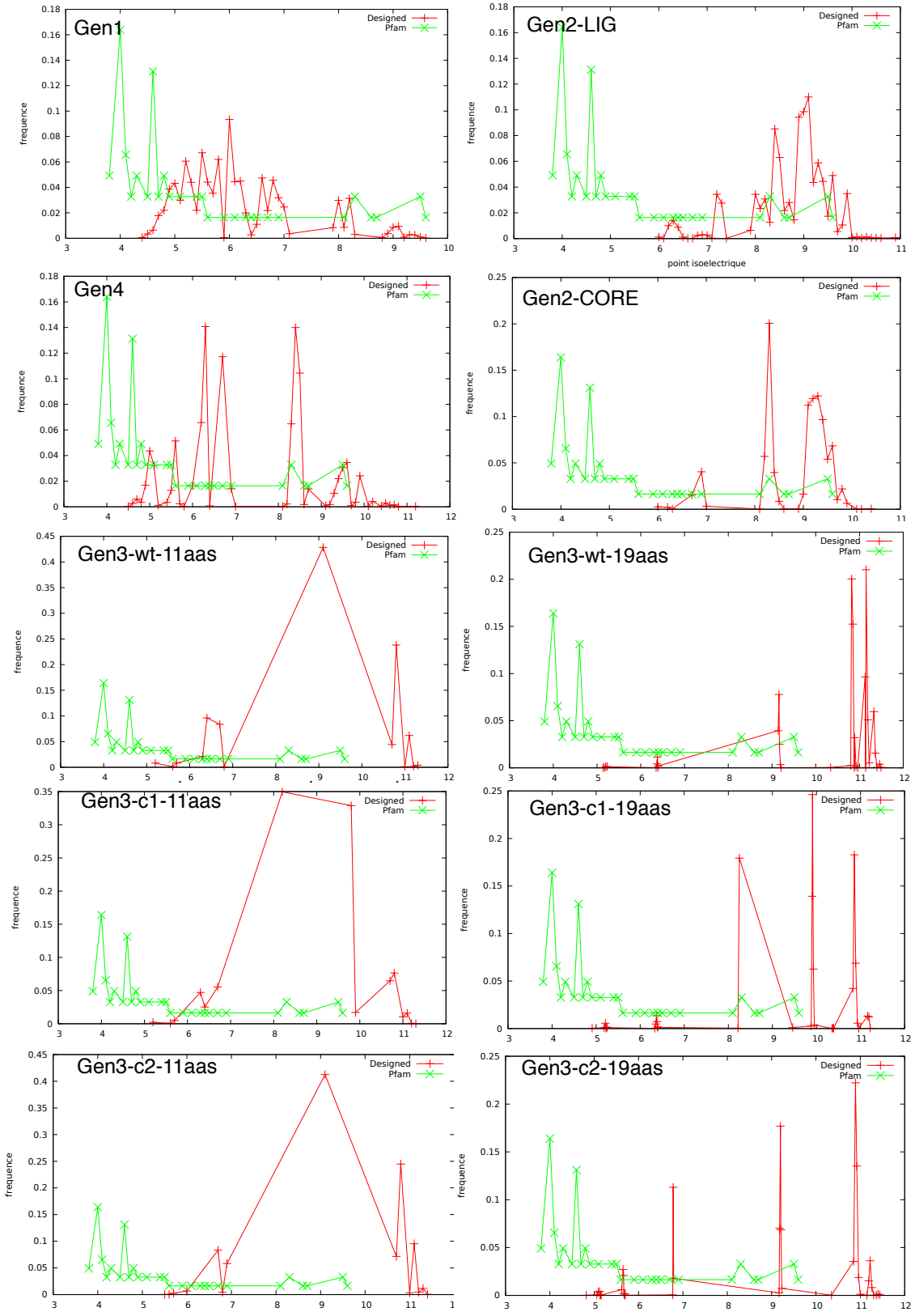


Figure 7.7 – Résultats des descripteurs : distribution des valeurs du point isoélectrique théorique pour les séquences naturelles PFAM et les séquences prédites (de Gen1 à Gen3).

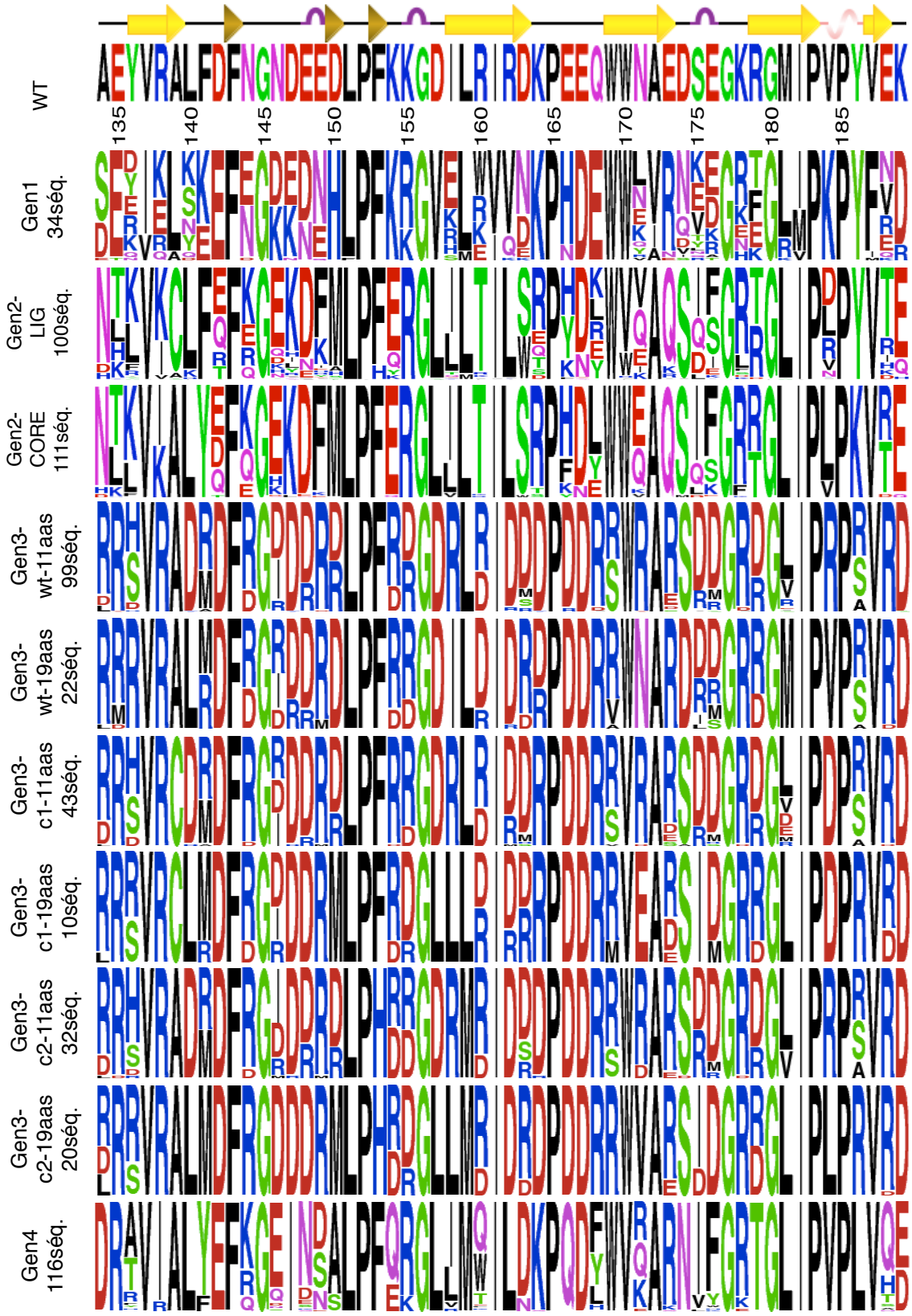


Figure 7.8 – Résultats des descripteurs : alignement des séquences sélectionnées par les filtres. Représentation logo (*weblogo*). Les acides aminés sont représentés par leur code à une lettre. Et la taille des lettres est proportionnelle à la fréquence d'apparition de l'acide aminé correspondant dans l'alignement de séquences considéré à la position courante.

Le tableau 7.3 résume l'impact sur le nombre de séquences considérées pour chaque génération, pour chaque filtre et chaque seuil choisi.

Génération séquences	Nb séquences de départ	Volume cœur	Simil		% Mut		pI
			seuil	nb seq	seuil	nb seq	~ 7
Gen1	5000	-	$\geq$47	264	$\leq$30	34	27
Gen2-LIG	5000	560	$\geq$34	189	$\leq$29	100	86
Gen2-CORE	5000	-	$\geq$55	591	$\leq$26	111	107
Gen3-wt-11	5000	-	$\geq$39	226	$\leq$29	99	83
Gen3-wt-19	5000	-	$\geq$46	333	$\leq$26	22	22
Gen3-c1-11	5000	-	$\geq$29	80	$\leq$28	43	43
Gen3-c1-19	5000	-	$\geq$24	111	$\leq$30	10	10
Gen3-c2-11	5000	-	$\geq$34	107	$\leq$30	32	27
Gen3-c2-19	5000	-	$\geq$28	168	$\leq$32	20	16
Gen4	5000	-	$\geq$51	415	$\leq$24	116	103

Table 7.3 – Résumé des résultats des filtres pour chaque génération. "Volume cœur" correspond au nombre de séquences ayant un volume du cœur hydrophobe compris entre 1094 et 1294 $/AA^3$. "Simil" est le nombre de séquences considérées (après le filtre précédent) ayant un score de similarité (blosum) contre l'alignement PFAM supérieur au seuil. "%Mut" correspond au nombre de séquences (après les 2 filtres précédents) ayant moins "seuil" % de mutations radicales par rapport à l'alignement PFAM. "pI" correspond au nombre de séquences (après les 3 filtres précédents) n'ayant pas un point isoélectrique théorique proche de 7.

7.2.1 Choix d'une séquence Gen1

La sélection de cette séquence théorique particulière (figure 7.9) a été beaucoup plus subjective et manuelle, et bien moins poussée. La simulation de dynamique moléculaire sur cette protéine mutante sera détaillée dans le chapitre suivant. Étudions d'abord à travers nos descripteurs, la qualité globale de l'ensemble de séquences prédites lors de la génération Gen1.

```
WT   AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMIPVPYVEK    4.56
     ...c.c.l.c.......c.c.....c.c.......lc.c.........c...lc..
old  DERVELSKEFEGDKEEHLPFKRGVKLRVQNKPHDEWWNVRQYDGKEGLVPKPYIND    5.63
```

Figure 7.9 – Séquence de la protéine mutante Gen1-*old* choisie pour l'étude expérimentale.

Nous avons pu déjà constater précédemment qu'il existe un déséquilibre dans l'abondance en acides aminés des séquences théoriques et celle dans la famille des domaines SH3 (trop de résidus chargés EKR, et pas assez de résidus IQSTA). Néanmoins, ce déséquilibre est moins important que celui de nos nouvelles générations de séquences théoriques. Cela peut s'observer directement dans les distributions des valeurs des points isoélectriques : ils sont plus proches de ceux de la famille SH3 dans la génération Gen1 que dans la génération Gen2. Les scores de similarité sont en moyenne de 41 alors que pour la génération "cœur fixe" précédente ils étaient d'environ 48 et pour la génération "positions fonctionnelles fixes" de 33.

7.2.2 Choix de séquences Gen2

Le schéma 7.10 décrit l'impact des différents filtres mis en place sur le nombre de séquences théoriques à considérer. Pour la génération de séquences théoriques Gen2 avec le cœur hydrophobe sauvage conservé, nous n'utilisons pas le filtre sur le volume du cœur hydrophobe puisque celui des protéines mutantes varient très peu (voir figure 7.4). La table 7.4 récapitule toutes les valeurs des descripteurs pour chacune des séquences

retenues, ainsi que pour trois séquences tests qui seront également testées en simulation de dynamique moléculaire afin de valider nos choix de seuils pour les filtres.

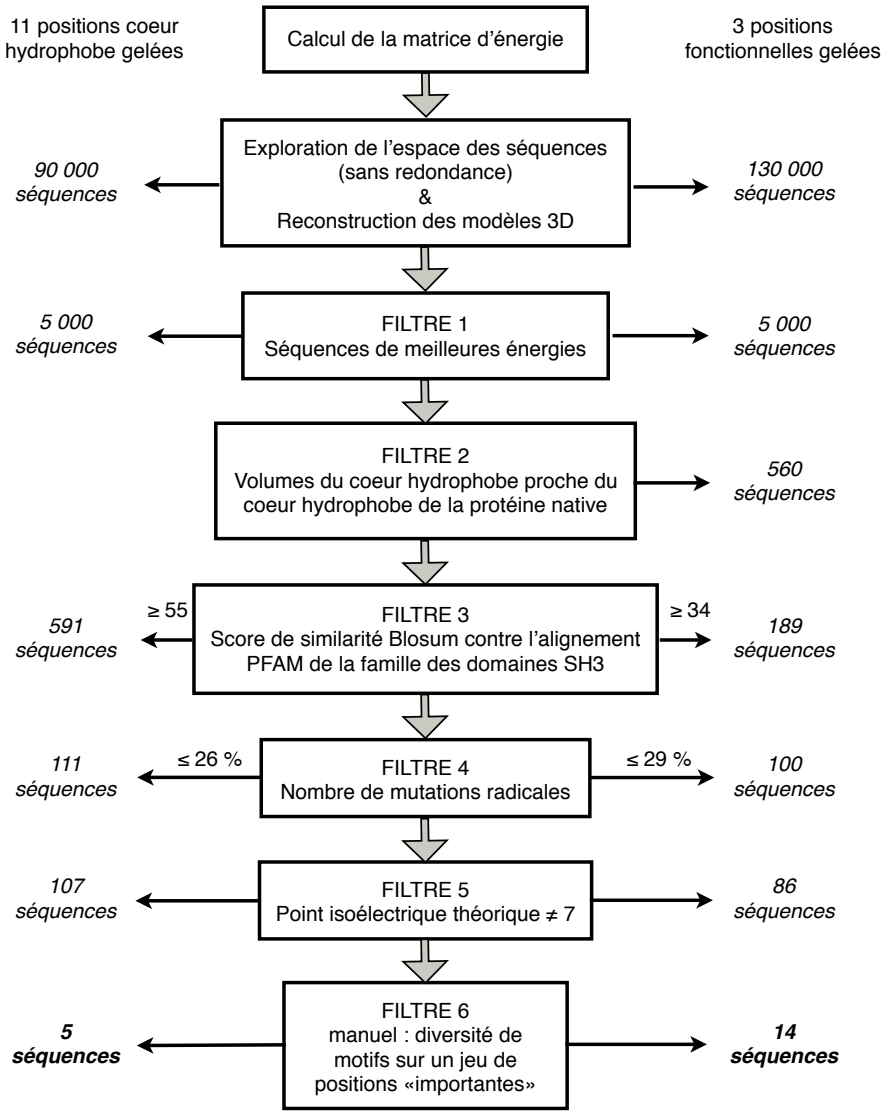


Figure 7.10 – Protocole de filtres pour sélectionner les séquences de la génération Gen2 à tester en expérimentale. On peut y voir pour les deux cas (LIG et CORE), l'impact des filtres sur le nombre de séquences considérées.

N_{seq}	E	Id_{wt}	S_{wt}	$\%_{wt\neq}$	S_{PFAM}	$\%_{PFAM\neq}$	Ch	$N_{\pm}$	$\%_{\pm}$	M	M/wt	M/PFAM	V	V/wt	V/PFAM	N_i	$\%_i$	V_c	pI
1CKA-WT																			
WT	-	100	307	0%	75.80	21%	-5	0	0%	6642	1.00	1.05	4994.50	1.00	1.06	0	0%	1194	4.56
Gen1-*old*																			
13	-	42.86	147	16%	47.52	32%	0	20	36%	6779	1.02	1.07	5161	1.03	1.10	5	9%	-	5.63
Gen2-LIG-V																			
10	249.70	37.50	109	18%	40.62	29%	2	21	38%	6496	0.98	1.03	4994.32	1.00	1.07	6	11%	1125.59	8.16
99	249.13	41.07	124	16%	46.66	27%	3	18	32%	6557	0.99	1.04	5076.42	1.02	1.09	6	11%	1116.01	9.36
114	249.11	42.86	114	12%	35.49	27%	2	19	34%	6661	1.00	1.06	5229.29	1.05	1.12	10	18%	1124.08	8.16
135	249.07	39.29	103	18%	34.46	26%	2	19	34%	6637	1.00	1.05	5197.61	1.04	1.12	9	16%	1138.94	8.15
433	248.57	42.86	121	18%	43.43	29%	3	21	38%	6456	0.98	1.03	4950	0.99	1.06	6	11%	1120	9.30
1445	247.84	39.29	103	20%	37.87	29%	3	25	45%	6515	0.98	1.03	5027	1.01	1.08	9	16%	1131	9.51
2857	247.15	39.29	105	20%	40.41	27%	2	21	38%	6466	0.98	1.02	4969	1.00	1.06	8	14%	1111	8.18
Gen2-LIG-W																			
105	249.12	39.29	117	20%	39.85	29%	3	20	36%	6714	1.01	1.07	5291	1.06	1.14	10	18%	1101	9.23
144	249.05	41.07	114	20%	39.97	29%	4	22	39%	6743	1.01	1.07	5327	1.07	1.15	12	21%	1109	9.60
421	248.58	39.29	118	16%	38.79	28%	4	20	36%	6731	1.01	1.07	5271	1.06	1.14	10	18%	1126	9.04
466	248.55	42.86	123	14%	40.72	29%	3	20	36%	6726	1.01	1.07	5313	1.06	1.14	11	20%	1213	8.94
589	248.43	39.29	120	20%	43.64	29%	3	21	38%	6543	0.99	1.04	5057	1.01	1.08	8	14%	1112	8.14
847	248.22	42.86	125	18%	43.67	28%	3	20	36%	6603	0.99	1.05	5138	1.03	1.11	8	14%	1213	9.05
1876	247.61	44.64	127	14%	42.07	29%	3	21	38%	6723	1.01	1.07	5288	1.06	1.14	11	20%	1106	9.16
Gen2-CORE																			
34	254.18	42.86	130	18%	59.38	25%	3	20	36%	6572	0.99	1.05	5101	1.02	1.11	8	14%	1201	9.30
245	253.70	39.29	118	20%	55.34	26%	3	20	36%	6529	0.99	1.04	5056	1.01	1.10	8	14%	1213	9.40
428	253.52	41.07	122	20%	56.18	26%	2	20	36%	6535	0.99	1.04	5069	1.01	1.10	9	16%	1219	8.39
1889	252.87	44.64	137	18%	58.84	26%	3	19	34%	6539	0.99	1.04	5061	1.01	1.10	7	12%	1199	9.23
2469	252.70	44.64	134	11%	55.75	25%	2	18	32%	6787	1.02	1.07	5388	1.08	1.16	10	18%	1247	8.34
Gen2-LIG "tests"																			
1962	-	33.93	94	11%	18.98	35%	5	26	46%	6723	1.02	1.08	5339	1.07	1.14	12	21%	1863	9.16
2533	-	39.29	107	14%	30.02	33%	4	25	45%	6685	1.00	1.05	5115	1.02	1.09	9	16%	907	9.30
4333	-	42.86	126	18%	49.84	25%	3	20	36%	6454	0.99	1.04	5163	1.03	1.10	8	14%	1246	8.34

Table 7.4 – Résumé des valeurs des descripteurs pour la séquence sauvage 1CKA-WT, la séquence choisie dans la génération Gen1-*old*, les séquences mutantes Gen2-LIG, Gen2-CORE, et 3 mutants Gen2-LIG tests. *De gauche à droite :* Numéro de séquence, énergie "PROTEUS", score identité par rapport à WT, score similarité par rapport à WT, % de positions très différentes par rapport à WT, score similarité par rapport à l'alignement PFAM des domaines SH3, % de positions très différentes par rapport à PFAM, charge, nombre de mutations de charge par rapport à WT, masse totale de la protéine, rapport masse / masse WT, rapport masse / masse moyenne PFAM, volume total de la protéine, rapport volume total / volume total WT, nombre de mutations ioniques par rapport à WT, % de positions avec mutation ionique par rapport à PFAM, volume du cœur hydrophobe, point isoélectrique.

Dans la dernière étape de sélection des séquences candidates, nous nous focalisons sur la position 170, et plus particulièrement sur la mutation W170V. Ainsi, nous regroupons nos séquences candidates en deux groupes : les séquences qui conservent le W_{170} natif, et les séquences qui ont la mutation V_{170}. Effectivement, muter un Tryptophane (W) en Valine (V) est un changement radical en terme de taille. Nous sélectionnons finalement 7 séquences mutantes avec le W_{170}, et 7 séquences mutantes avec la mutation V_{170}, afin de les tester en dynamique moléculaire (voir la figure 7.11).

```
Wt    AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMIPVPYVEK 4.56
      ...c.c.l.c.......c.c.....c.c.......lc.c.........c...lc..
10    NTLVKCLFTFRGEKDFMLPFERGLLLTILSRPHDKWVQAQSIFGRTGLIPDPYVTE 8.16
99    NTLVKCLFEFEGQKDKMLPFKRGLLLTILSRPYDKWVEAQSIFGRRGLIPDPYVTQ 9.36
114   NTLVICLFEFKGNLERMLPFERGLLLTILWRPYDKWVVAQSIFGRRGLIPDPYVHE 8.16
135   NTKVKCLFEFKGEKDFMLPFERGLILTILWRPHDLWVVAQSIFGRTGLIPWPYVTE 8.15
433   NTKVKCLFTFRGEKDFMLPFERGLILTILSRPKDKWVQAKSLEGKTGLIPDPYVTE 9.30
1445  NTKVKCLFTFRGEKNFMLPFERGLILTILSRPYNEWVQAQSQIGLTGLIPRPYVTE 9.51
2857  NTRVVCLFQFKGEKDFMLPFERGLTLTILSRPHDLWVQAQSIFGRTGLIPVPYVTE 8.18
      ...c.c.l.c.......c.c.....c.c.......lc.c.........c...lc..
105   NKLVKALFEFKGEKDFMLPHERGLLMTILWTPYNYWWVAQSIFGRRGLIPLPYVTE 9.23
144   DLKVKALFTFEGKKDFMLPHERGLIMTILWRPYNYWWVAQSIFGRRGLIPLPYVTE 9.60
421   NLLVKCLFQFEGHKDFMLPHERGLLLTILWRPYNYWWVAQSQSGRRGLIPLPYVTE 9.04
466   NLKVKCLFEFKGEKDFMLPFYRGLILTILWRPHDLWWVAQSDKGFRGLIPLPYVTE 8.94
589   NKKVKCLFEFKGEKDFMLPHERGLILTILSTPHDLWWQAQSIFGRTGLIPLPYVTE 8.14
847   NTKVKCLFQFKGEKDFMLPFQRGLILTILWRPHDLWWVAQSLSGDRGLIPLPYVTE 9.05
1876  DLKVKALFQFKGEYDFMLPHERGLIMTILWRPKNYWWVAQSLEGKRGLIPLPYVTE 9.16
```

Figure 7.11 – Alignement des séquences théoriques Gen2-LIG mutantes choisies pour la simulation en dynamique moléculaire. Les positions du cœur hydrophobe sont marquées d'un "c", et les positions fonctionnelles d'un "l".

Nous avons également choisi trois séquences mutantes qui n'avaient pas été sélectionnées par nos filtres (figure 7.12). Ces protéines mutantes seront des références pour évaluer l'efficacité de nos sélections par les descripteurs et la pertinence des seuils choisis. A partir de la génération Gen2-LIG, nous avons sélectionné ces mutants pour vérifier la solidité de nos seuils (1094-1294 $Å^3$) pour le filtre sur le volume du cœur hydrophobe : 1CKA-LIG-1962 (1863 $Å^3$), 1CKA-LIG-4333 (1246 $Å^3$) et 1CKA-LIG-2533 (907 $Å^3$). Les valeurs des

descripteurs sont répertoriées dans le tableau 7.4. Le cas du mutant 1CKA-LIG-4333 est particulier : il respecte tout les filtres choisis, et semble être très ressemblant en terme de séquence à la protéine sauvage ainsi qu'à la famille des domaines SH3. En effet, le volume de son cœur hydrophobe est en dessous du seuil maximum (1294 $Å^3$), son score de similarité avec la famille PFAM (49.89) respecte largement du seuil choisi (>34), et son pourcentage de mutations très différentes (25%) respecte également le seuil (<29%), mais cette protéine possède néanmoins le plus volumineux cœur des séquences sélectionnées par ces filtres. Les simulations de dynamique moléculaire sur ce mutant nous confirmerons ou non le bon choix du seuil supérieur pour le volume du cœur hydrophobe. Pour les deux autres séquences mutantes, leur volume du cœur est nettement en dehors des valeurs acceptées (bien plus petit et bien plus grand), et en terme de séquences elles sont très éloignées de la protéine sauvage et de la famille des domaines SH3. Les simulations en dynamique sur ces deux protéines mutantes devraient confirmer ces observations par des comportements moins bons, et donc valider le choix de nos seuils.

```
WT      AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMIPVPYVEK
        ...c.c.l.c.......c.c.....c.c.......lc.c.........c...lc..
1962    HTFVICRFQFRGEYDKMLPHYRGLVLWIKWEPKDRWVVAKSDKGFTGLIPRPYVDE
2533    HTKVKALFTFRGEKDKMLPHERGLIMWIKWEPHDEWVVAESDKGFTGLIPRPYVTQ
4333    NTKVICLFEFKGEKEIHLPFERGLILTILSRPHDLWWIAQSIFGRTGLIPLPYVRE
```

Figure 7.12 – Alignement des séquences théoriques tests non sélectionnées par nos filtres pour la simulation en dynamique moléculaire. Les positions du cœur hydrophobe sont marquées d'un "c", et les positions fonctionnelles d'un "l".

De la même manière, nous pouvons justifier les choix de seuil pour les descripteurs "volume du cœur hydrophobe", "score Blosum contre PFAM" et "nombre de mutations radicales" pour les séquences de la génération Gen2-CORE.

Dans la dernière étape de sélection des séquences candidates, nous choisissons des séquences dont les positions 142, 149, 171 et 176 conservent l'acide aminé natif ou la charge négative du résidu natif, ainsi que certaines séquences avec ces 4 positions mutées plus radicalement. Nous sélectionnons finalement 5 séquences mutantes candidates pour

la simulation en dynamique moléculaire, ayant une variabilité représentative sur les 4 positions précédentes (voir la figure 7.13).

```
Wt    AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMIPVPYVEK  4.56
      ...c.c.l.c.......c.c.....c.c.......lc.c.........c....lc..
34    NTKVKALYDFKGEKDFMLPFERGLILTILSRPHDLWWEAQSIFGRRGLIPLPKVTE  9.31
245   NTKVKALYQFQGEKDFMLPFERGLILTILSRPHDLWWQAQSIFGRTGLIPLPKVTE  9.40
428   NTKVKALYDFEGHKDFMLPFERGLILTILSRPFDLWWQAQSIFGRTGLIPLPKVTE  8.39
1889  NTKVKALYEFKGEKDFMLPFQRGLILTILSRPHDLWWEAQSLEGKRGLIPLPKVTE  9.23
2469  DLFVIALYEFKGEKDEMLPFERGLILRILWRPKDYWWVAQSIFGRRGLIPLPKVHE  8.34
```

Figure 7.13 – Alignement des séquences théoriques Gen2-CORE choisies pour la simulation en dynamique moléculaire. Les positions du cœur hydrophobe sont marquées d'un "c", et les positions fonctionnelles d'un "l".

7.3 Travaux sur une autre structure

Nous avons appliqué nos descripteurs sur des séquences théoriques calculées à partir d'une autre structure d'un domaine SH3 : 1CSK. Les analyses sont en annexes. Ces séquences ont été générées avec les paramètres de la génération Gen2. On peut observer que ces séquences ont des scores de similarité désastreux.

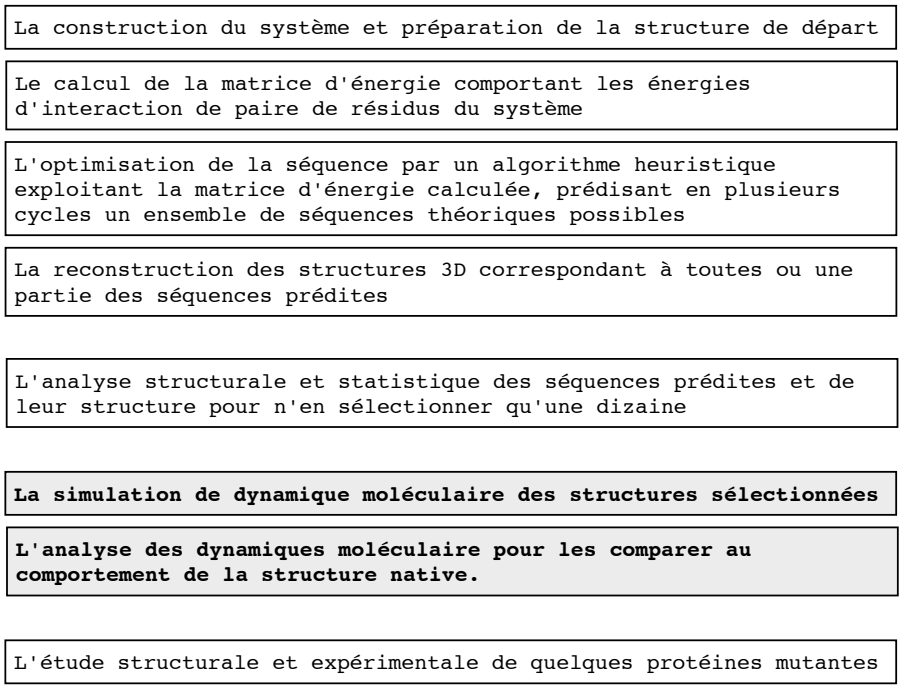
7.4 Conclusion

Cette analyse bioinformatique mesure la qualité des séquences prédites par CPD. Nous pouvons ainsi identifier les différentes générations prédisant les séquences théoriques les plus représentatives et les plus proches de la famille des domaines SH3 et de la structure particulière SH3-1CKA. Les descripteurs choisis paraissent être de bons indicateurs sur la ressemblance avec les séquences naturelles et nous ont permis de sélectionner 19 séquences candidates qui, d'un point de vue purement théorique, semblent avoir des caractéristiques nettement améliorées par rapport à la séquence choisie Gen1-*old*.

Avant de vérifier la structure expérimentale de ces protéines mutantes, il convient cependant d'étudier leur stabilité *in silico* par des simulations de dynamique moléculaire. C'est que nous allons détaillé dans le chapitre suivant.

Chapitre 8

Étude par simulation de dynamique moléculaire des séquences théoriques



Dans ce chapitre, nous décrirons plus précisément la suite du protocole de sélection des protéines les plus prometteuses (en gras ci-dessus). Ainsi, après la génération de séquences théoriques, l'exploration de l'espace des séquences possibles et la reconstruction de leur structure, puis la caractérisation et la sélection des séquences les plus pertinentes par le biais de descripteurs, nous procédons à une simulation de dynamique moléculaire qui

servira à tester la stabilité *in silico* de l'ensemble de séquences théoriques sélectionnées dans le chapitre précédent.

Nous simulons ici les séquences théoriques des générations Gen2 choisies dans le chapitre précédent. Afin de juger de la stabilité de nos protéines mutantes, nous comparerons l'analyse de leur simulation avec la protéine mutante *old* de la génération Gen1.

8.1 Simulation de dynamique moléculaire

Jusqu'ici les structures étaient assez figées et entourées de solvant implicite. C'est pourquoi nous testons nos prédictions structurales à l'issue de la méthode de CPD au cours de simulations de dynamique moléculaire avec solvant explicite. La dynamique moléculaire consiste à calculer l'évolution d'un système de particules au cours du temps. Elle requiert une fonction d'énergie capable de décrire les forces qui guideront les différents atomes de la protéine durant la simulation et s'appuie principalement sur des potentiels physiques connus sous le nom de "champ de force".

Les dynamiques moléculaires que nous réalisons dans notre équipe, utilisent le programme CHARMM [Brooks *et al.* 1983] ainsi que le champ de force CHARMM22. Nous appliquons l'algorithme d'intégration de Verlet (décrit dans le chapitre "Introduction au CPD") avec un pas d'intégration de 1 fs. Le système est d'abord translaté et réorienté pour aligner son centre de masse et ses axes principaux d'inertie avec le système de coordonnées cartésiennes. Il est ensuite solvaté avec une boîte parallélépipédique de molécules d'eau explicites TIP3P, préalablement équilibrées à une pression d'une atmosphère et une densité de 0,0334 molécules par $Å^3$. Des contre-ions (chloride ou sodium) sont ajoutés pour neutraliser le système, par remplacement de molécules d'eau. Nous appliquons des contraintes structurales autour de la boîte d'eau, afin de conserver la compacité de l'ensemble. On adopte des conditions aux limites périodiques, en considérant un réseau de cubes images par symétrie de translation. Les dimensions de la boîte ont été choisies pour

dépasser la protéine d'au moins 10 Å dans toutes les directions, afin d'être suffisamment grande pour éviter des artéfacts de périodicité.

La structure est minimisée pendant 100 pas dans sa boîte d'eau, où nous maintenons constante la température du système par un thermostat de type Nosé-Hoover [Spiegel *et al.* 2006]. Et les interactions électrostatiques à grandes distance sont traitées avec la méthode "Particule Mesh Ewald" [Tozzini 2005]. Nous effectuons des simulations de 20 ns à température ambiante et constante (295 K) pour refléter une température physiologique. Nous avons utilisé les ressources du super calculateur du CINES (Centre Informatique Nationale de l'Education Supérieure) pour faire nos simulations de dynamique moléculaire.

8.1.1 Analyse des simulations de dynamique moléculaire

Il existe plusieurs grandeurs pertinentes pour décrire des systèmes biologiques, néanmoins le rayon de giration, le RMSD et le volume structural seront les observables privilégiés dans le cadre de cette analyse des dynamiques moléculaires.

8.1.1.1 Calcul du rayon de giration

Le rayon de giration donne une mesure de la compacité du système et peut donc être utilisé pour évaluer les changements structuraux de la protéine pendant la dynamique. Le rayon de giration est obtenu en moyennant les distances de chaque atome au centre de masse. Chaque contribution à la moyenne est pondérée en fonction de la masse de l'atome considéré. On le définit [Berne & Pecora 1976] comme :

$$Rg = \sqrt{\frac{\sum_i m_i (r_i - r_{CM})^2}{\sum_i m_i}} \tag{8.1}$$

où $r_{CM})^2$ est la position du centre de masse, r_i est la position de l'atome i et m_i sa masse.

8.1.1.2 Calcul du RMSD

Une deuxième analyse de l'évolution de la structure globale de la protéine à laquelle il est souvent fait appel, est l'écart quadratique moyen entre la structure au cours de la simulation et la structure de départ (RMSD). Le RMSD [Maiorov & Crippen 1994] prend en compte la distance scalaire entre atomes du même type comparant deux structures. Les mouvements de rotation et translation du système ont été supprimés des trajectoires par superposition. La superposition ainsi que le calcul du RMSD ont été effectués pour chaque structure sauvage et mutante étudiée afin d'évaluer les déformations au sein des protéines. Nous nous sommes limités aux RMSD des carbones α au cours du temps.

Pour deux ensembles de n point V et W à comparer, le RMSD est défini par :

$$RMSD_{(V,W)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2} \tag{8.2}$$

8.1.1.3 Calcul du volume structural du cœur hydrophobe

Les calculs des volumes structuraux des cœurs hydrophobes sont réalisés de la même manière (volume de Voronoï) que pour l'étape de filtration des séquences théoriques. Nous définissons deux ensembles de positions appartenant au cœur hydrophobe de la protéine. Le premier est l'ensemble des 11 positions que nous avons choisies (voir les chapitres précédents). Le deuxième est déterminé par le programme XPLOR, en sélectionnant les résidus enfouis. Nous obtenons généralement entre 15 et 18 acides aminés enfouis, représentant un cœur hydrophobe plus large. Nous pouvons ainsi observer les variations de volume au sein du cœur hydrophobe, témoignant donc directement de la stabilité de la protéine mutante.

8.2 Analyses des simulations de dynamique moléculaire sur la protéine SH3-1CKA

Afin de pouvoir appréhender les comportements des protéines mutantes lors des simulations de dynamique moléculaire, nous avons réalisé les mêmes analyses sur la simulation faite sur la structure et la séquence native 1CKA (Figure 8.1).

8.2.1 Protéine sauvage 1CKA-WT

Nous pouvons tout d'abord analyser les 20 ns de simulation sur la structure native. Le volume structural du cœur hydrophobe varie très peu (autour de 1194 $Å^3$), comme on peut le voir sur la figure 8.1. Le RMSD de la protéine entière varie entre 0,5 et 1 Å, mais reste principalement autour de 0,7 Å . Pour le cœur hydrophobe, le RMSD varie peu autour de 0,4 Å. Les rayons de giration de la protéine entière et du cœur hydrophobe sont également très stables, respectivement autour 9,75 et 6,95 Å. On ne suppose donc *a priori* aucun changement de volume, de compacité ou dépliement de la structure sauvage. La protéine sauvage "respire" principalement au niveau des boucles et un peu à l'extrémité N-terminale. Ces comportements joueront le rôle de référence pour comparer les dynamiques moléculaires des protéines mutantes.

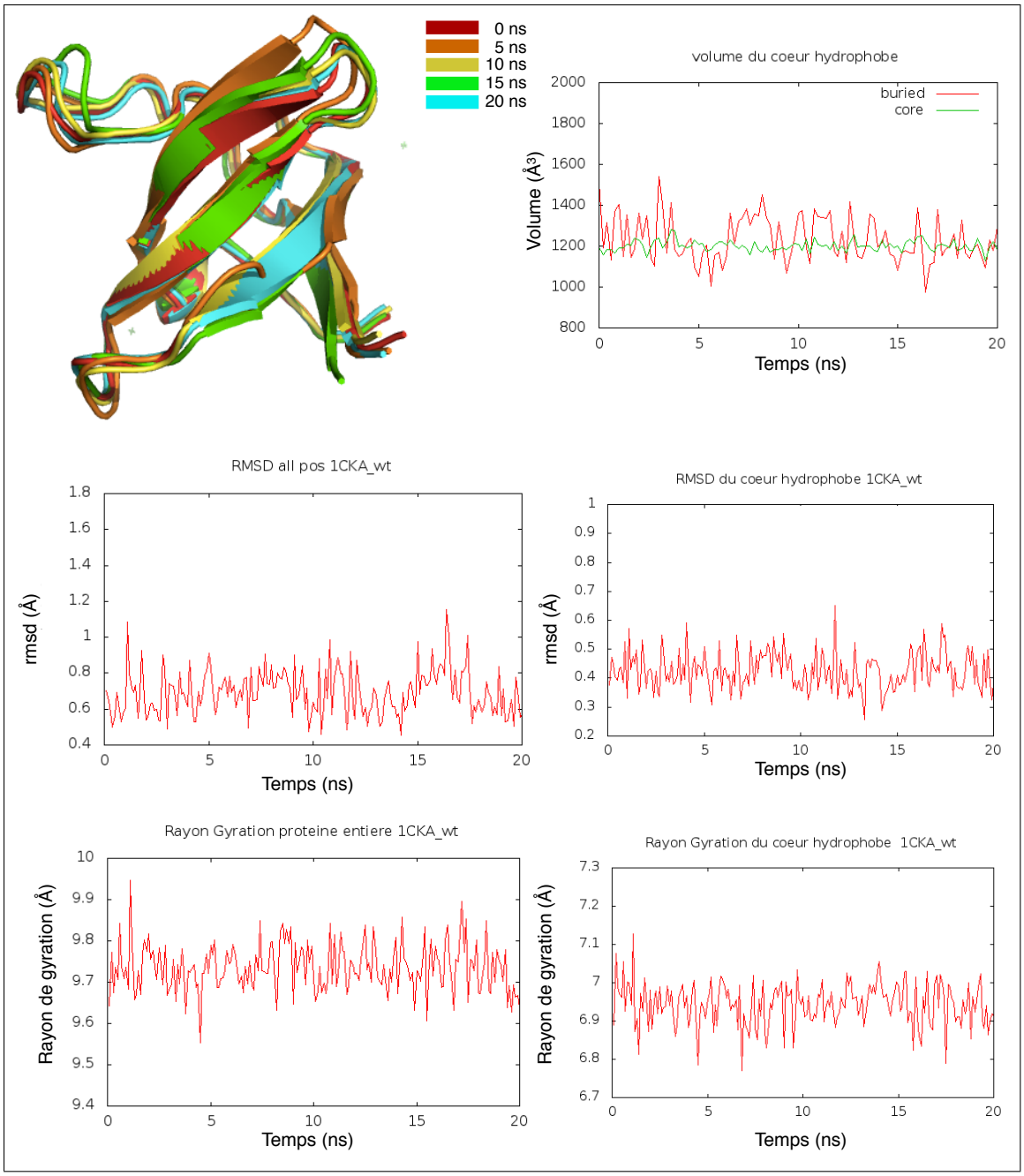


Figure 8.1 – Analyses de la simulation de dynamique moléculaire sur la structure native 1CKA : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.2.2 Protéines mutantes de la génération Gen2-LIG

Globalement, nous pouvons constater que les protéines mutantes de la génération Gen2-LIG sont assez stables : aucune ne se déplie complètement. Néanmoins, certaines "respirent" peut-être un peu plus. Les rayons de giration ne bougent pratiquement pas, ils restent autour de 9,7 Å pour la protéine entière. Pour le rayon de giration du cœur hydrophobe, il semble ne pas varier ou très peu, et rester autour du rayon de giration du cœur hydrophobe natif (7 Å).

8.2.2.1 Séquences théoriques avec la mutation V170

Prenons le cas des protéines mutantes avec les positions fonctionnelles fixées et la mutation V_{170} (Figures 8.2 à 8.8). Les protéines 10, 99 et 433 semblent moins bouger au niveau de l'extrémité N-terminale où se trouve la position 137 définie dans le cœur hydrophobe. Ainsi cela se témoigne par des valeurs de RMSD dans le cœur hydrophobe plus stables et moins élevées (autour de 0,6 Å). Alors que les protéines 114, 135, 1445 et 2857 semblent éloigner de plus en plus leur extrémité N-terminale vers l'extérieur (comme les protéines 114 et 135) ou vers l'intérieur (comme les protéines 1445 et 2857). Le RMSD global des protéines mutantes 10, 99, 114, 135, 433 et 1445 sont en moyenne de 1 Å, ce qui est déjà nettement meilleur que nos simulations précédentes (voir les analyses du mutant 1CKA-*old*). Et le RMSD du cœur hydrophobe pour ces mêmes protéines est en moyenne de 0,6 Å, ce qui n'est pas trop éloigné de celui de la protéine sauvage. Le cas de la protéine 2857 est différent : les RMSD de la protéine entière et du cœur hydrophobe semble augmenter au fur et à mesure de la dynamique moléculaire. Cependant, il faudrait plus de 20 ns pour tirer une conclusion sur une déviation notable et la stabilité réelle de ces protéines mutantes.

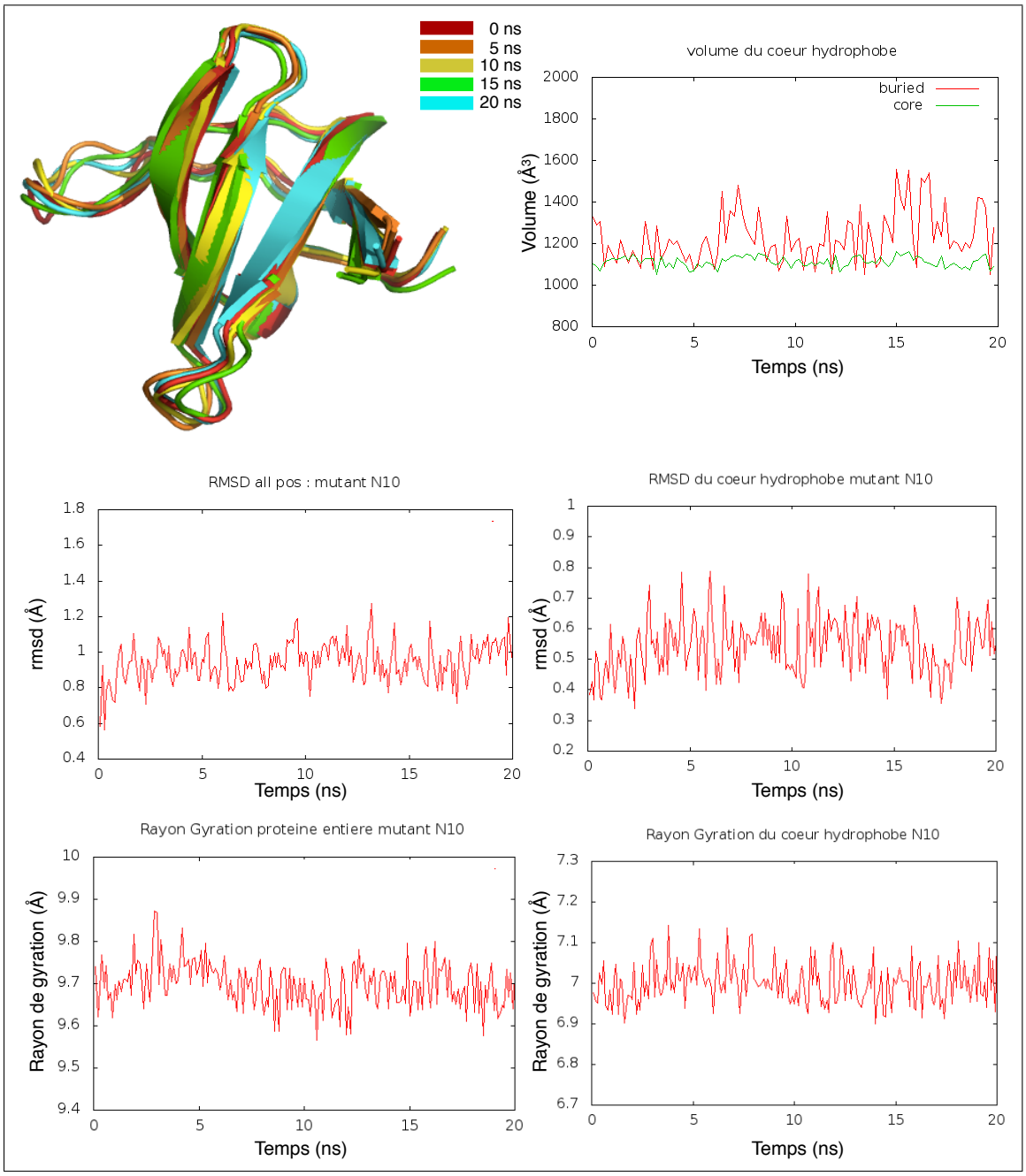


Figure 8.2 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N10 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

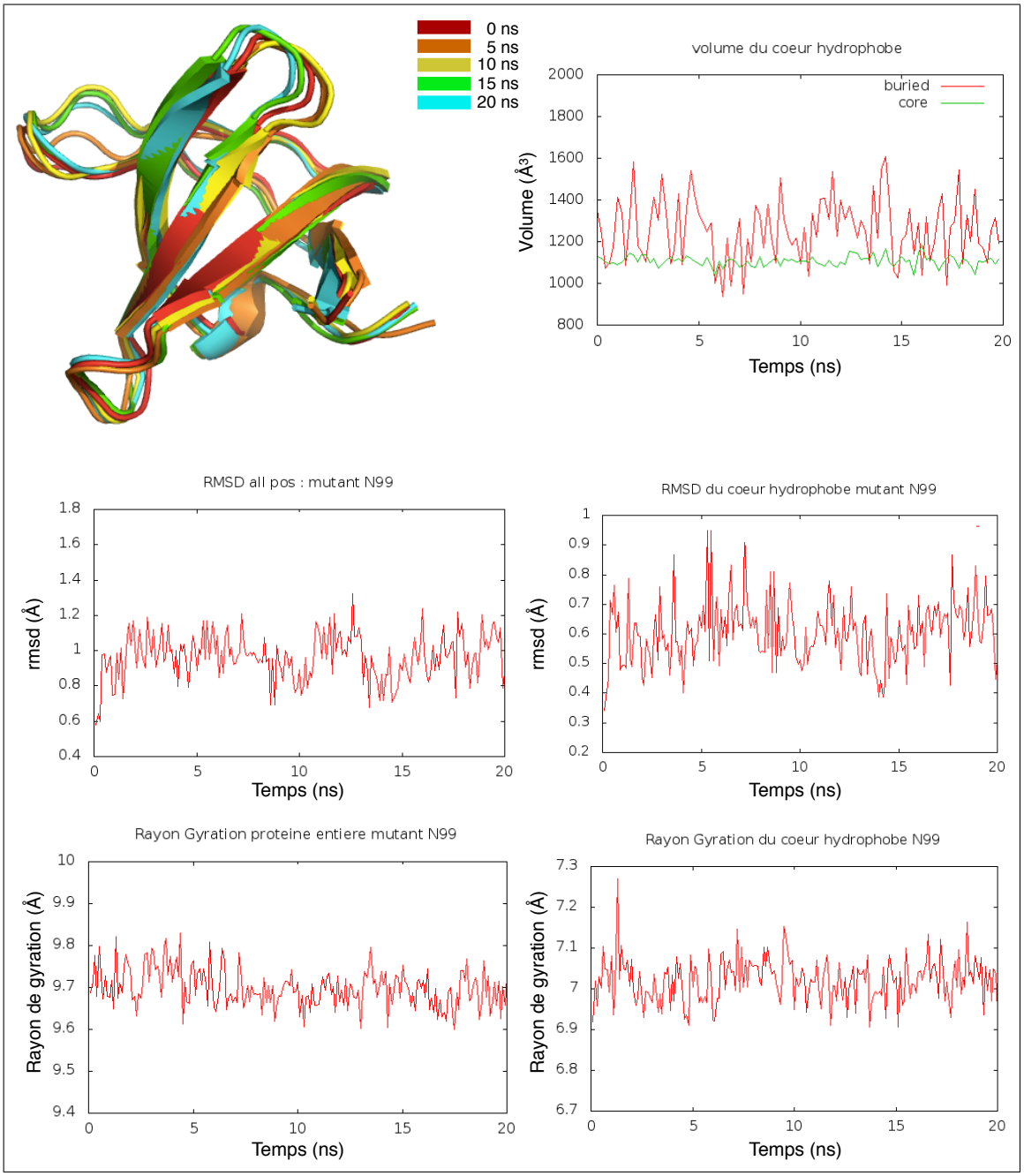


Figure 8.3 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N99 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

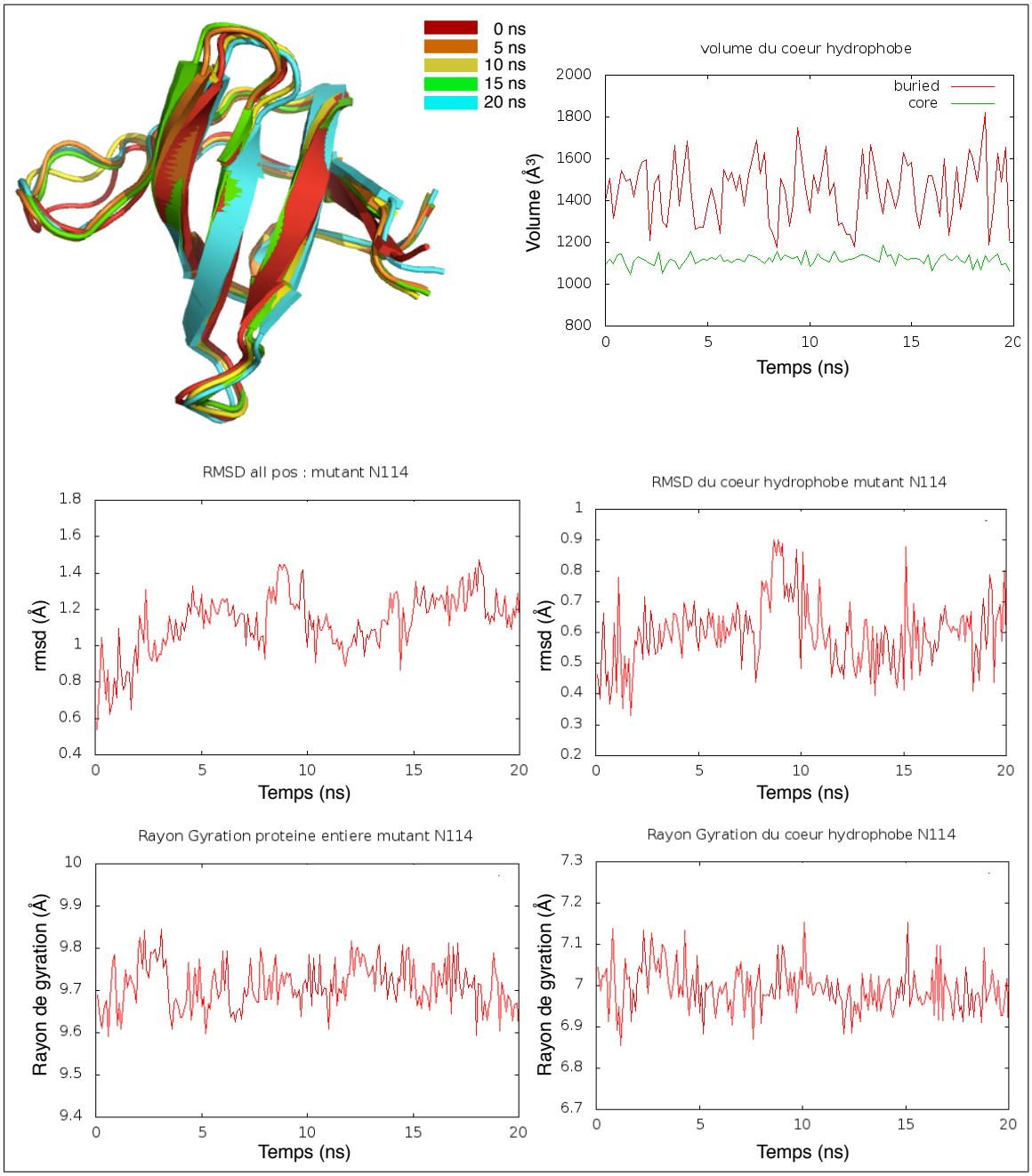


Figure 8.4 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N114 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

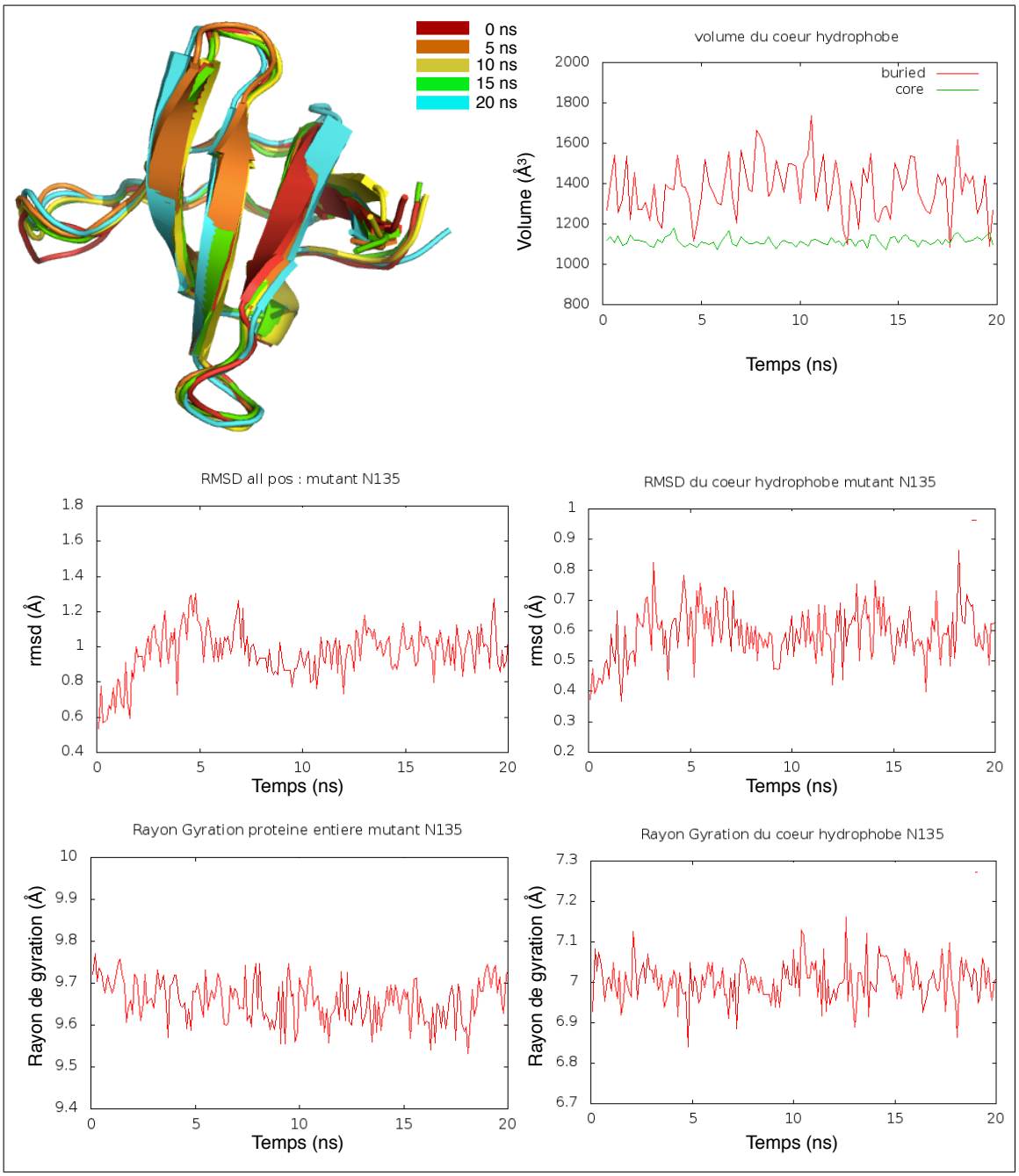


Figure 8.5 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N135 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

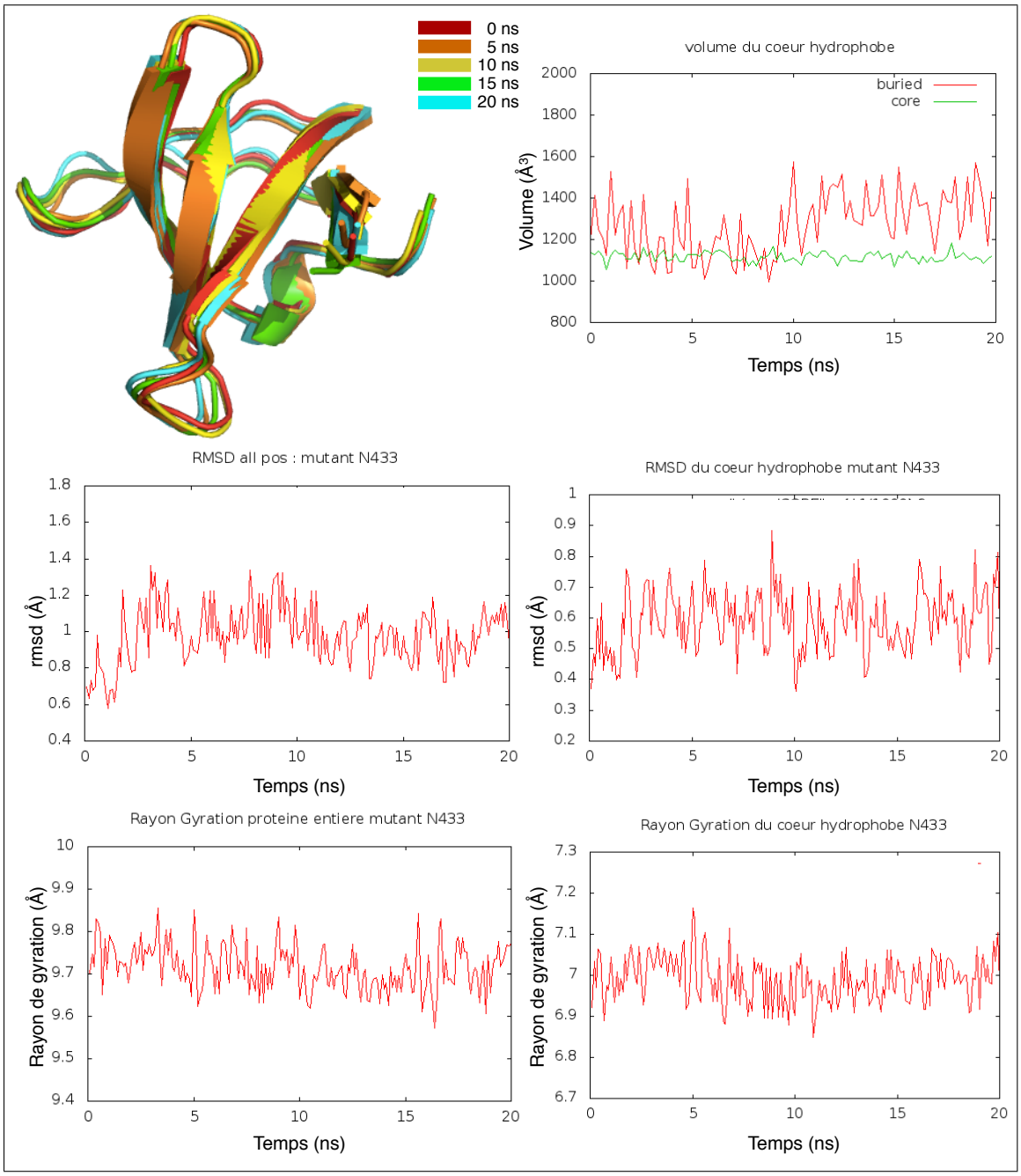


Figure 8.6 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N433 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

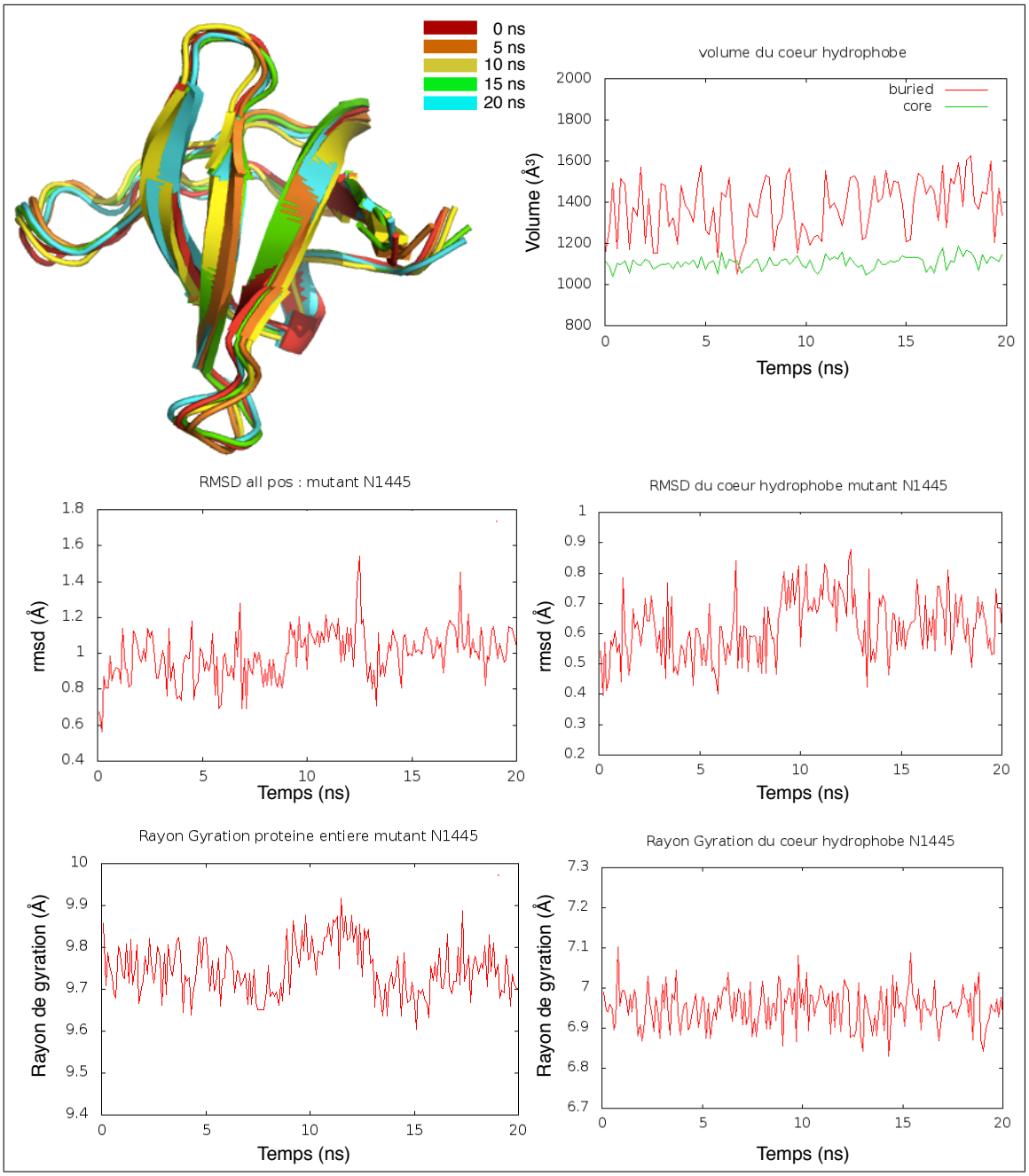


Figure 8.7 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N1445 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

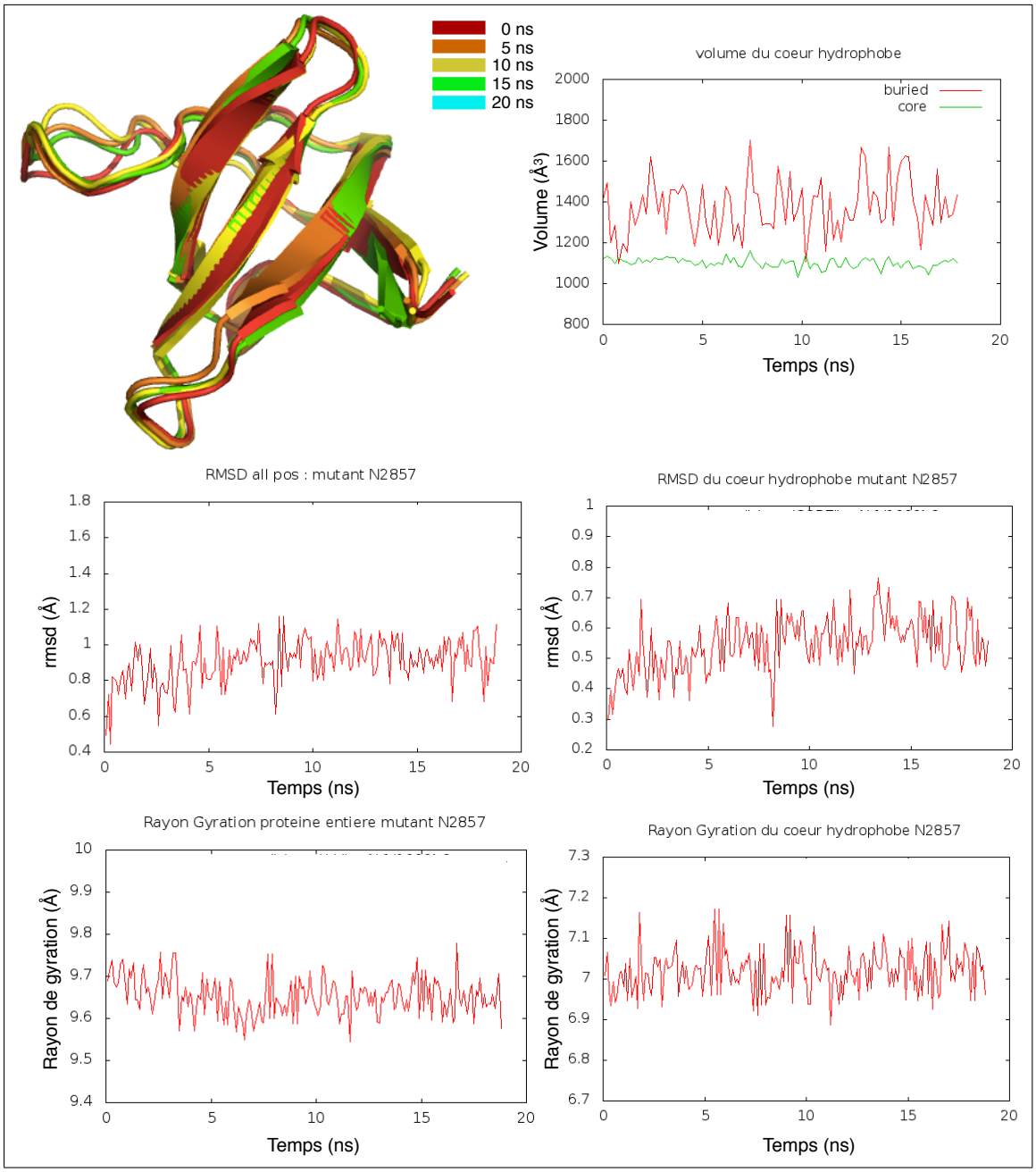


Figure 8.8 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N2857 de la génération Gen2-LIG avec la mutation V_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.2.2.2 Séquences théoriques sans mutation en position W170

Si on observe les protéines mutantes de la génération Gen2-LIG sans mutation en W_{170} (Figures 8.9 à 8.14 en annexe), on peut voir que leur cœur hydrophobe bouge moins, et donc que ces comportements se rapprochent plus de celui de la protéine sauvage. En effet, les RMSD du cœur hydrophobe de ces sept protéines mutantes restent autour de 0,5 Å. Le RMSD global est entre 0,8 et 1,1 Å pour ces protéines. En revanche, les extrémités N-terminales des protéines 847 et 1876 semblent ne pas bouger, contrairement aux protéines 105 et 589.

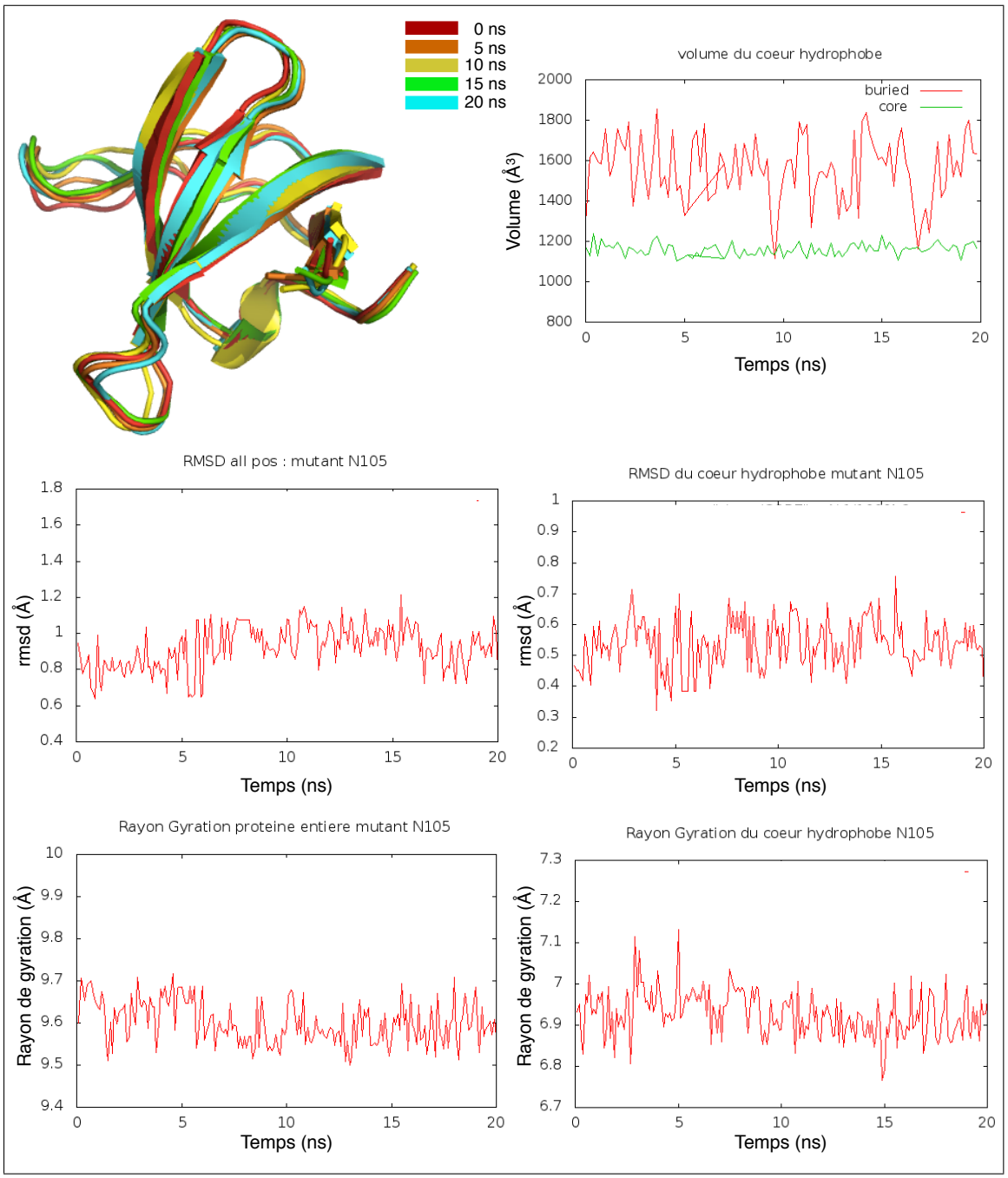


Figure 8.9 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N105 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

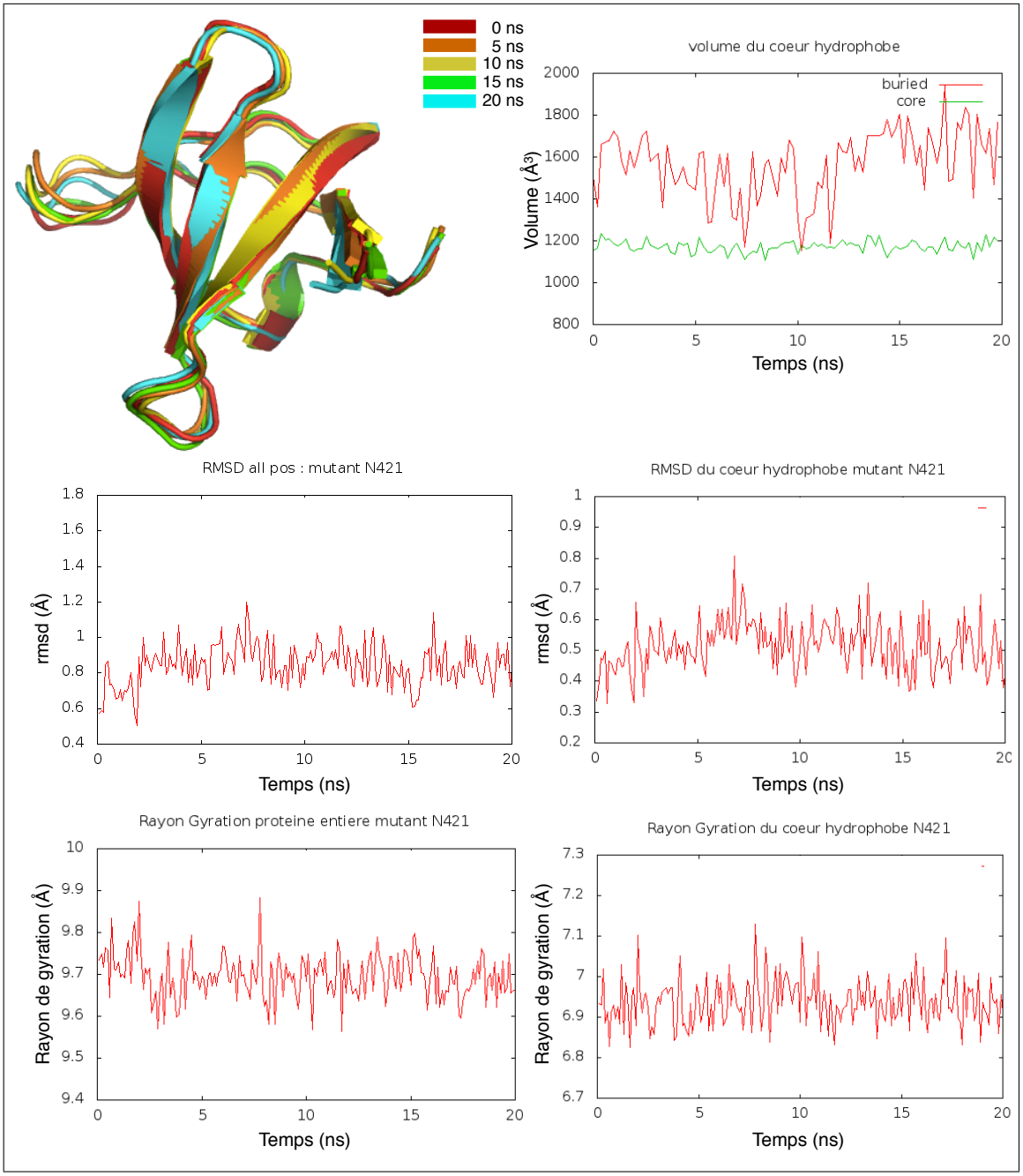


Figure 8.10 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N421 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

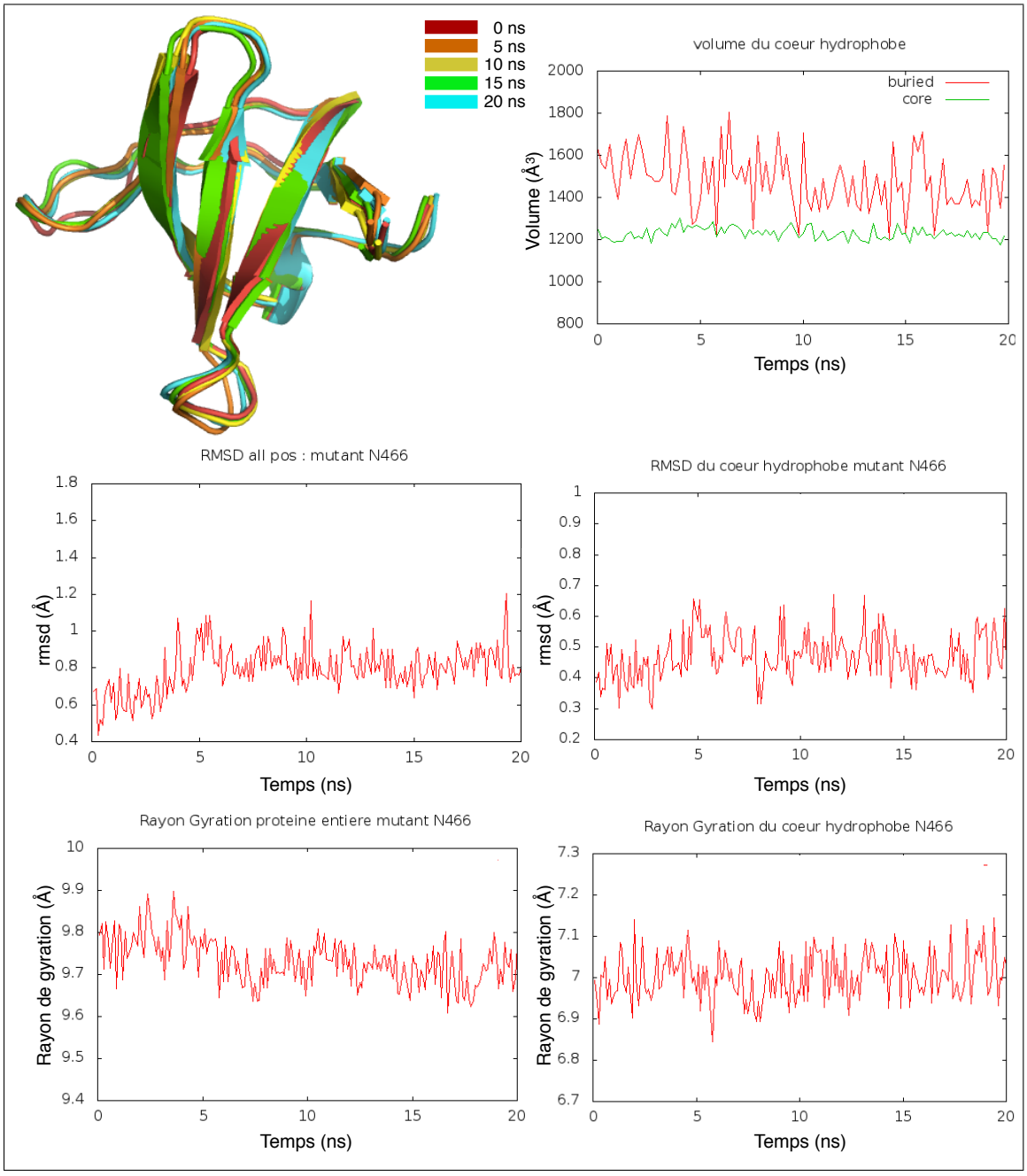


Figure 8.11 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N466 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

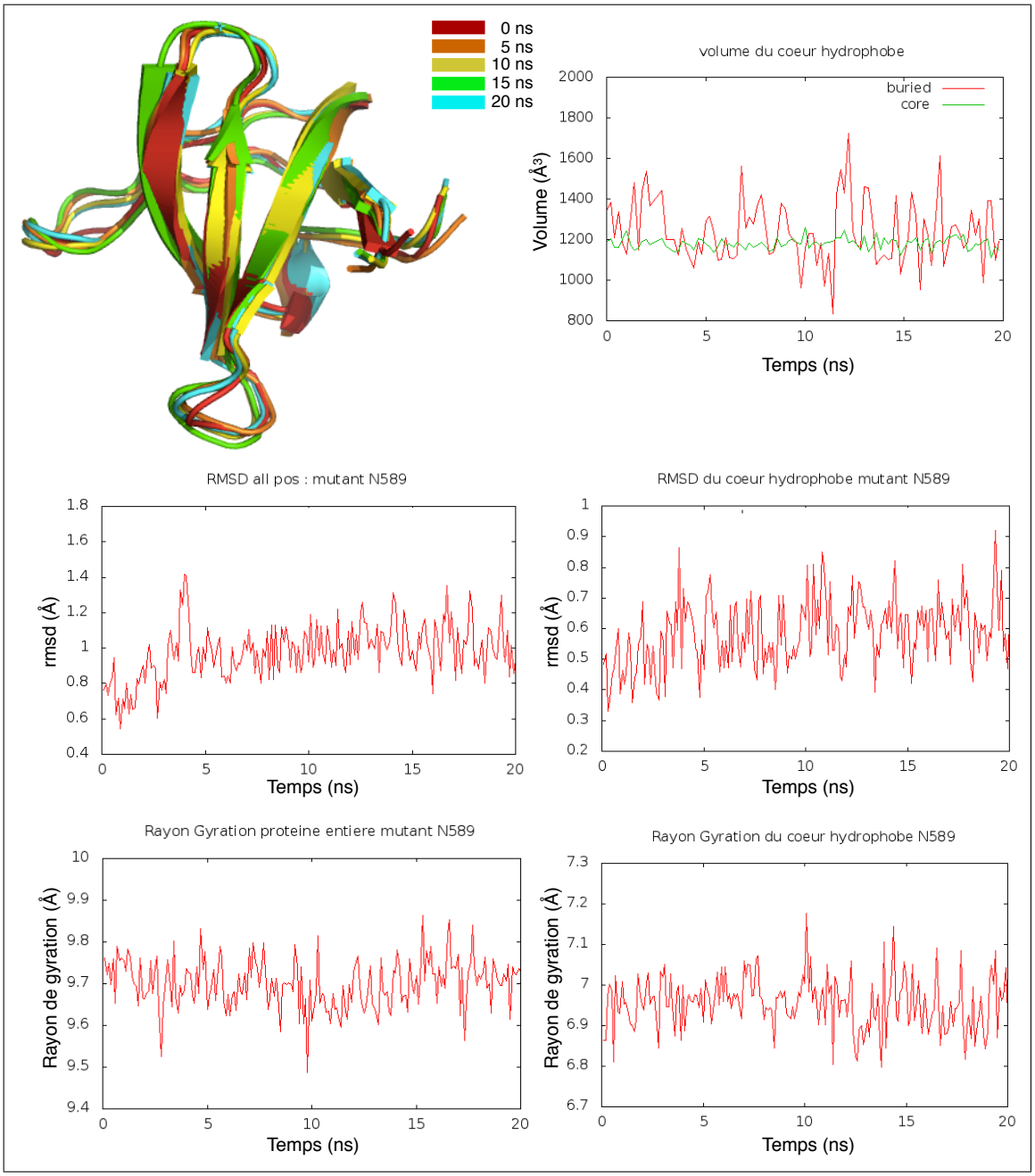


Figure 8.12 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N589 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

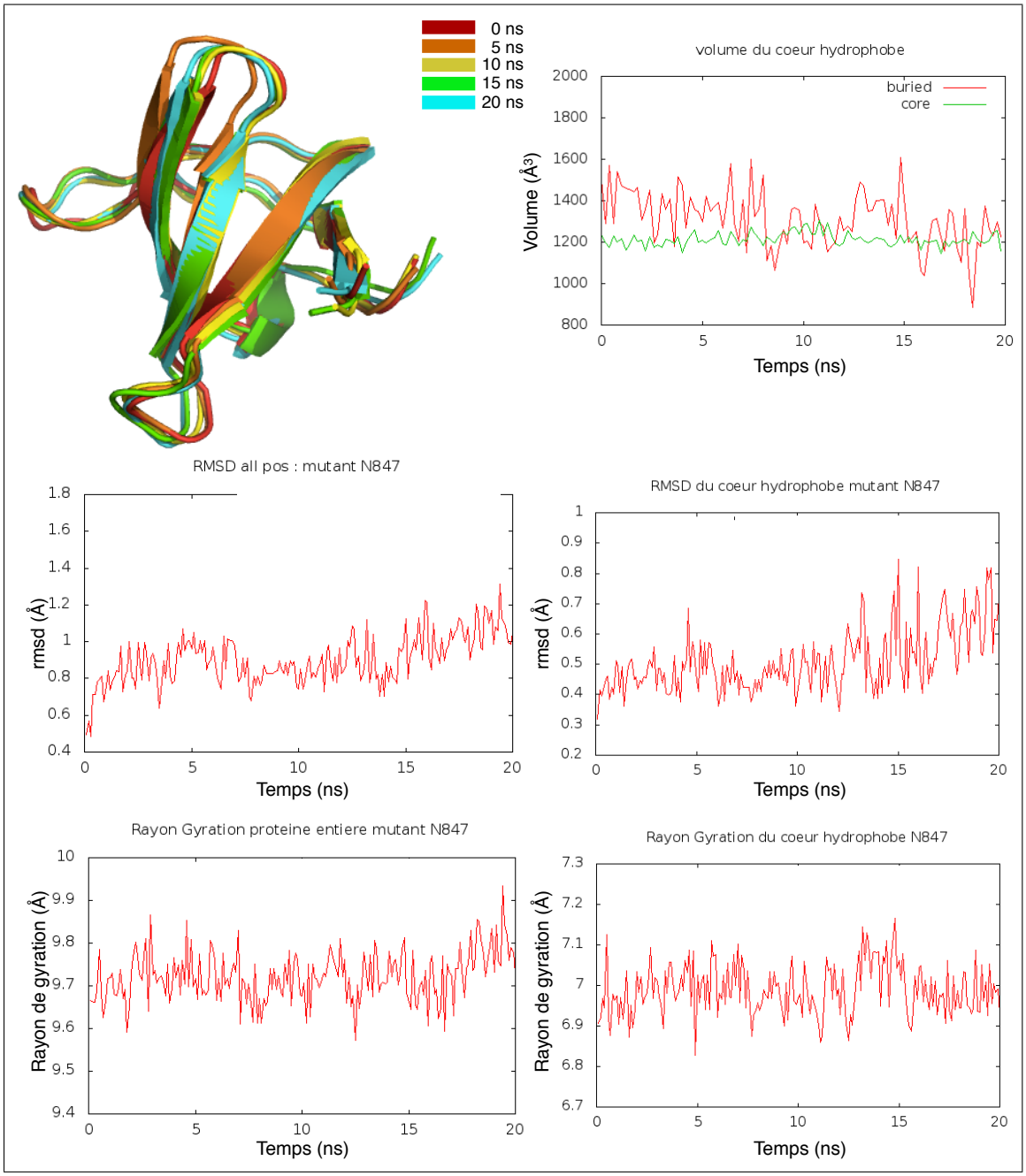


Figure 8.13 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N847 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

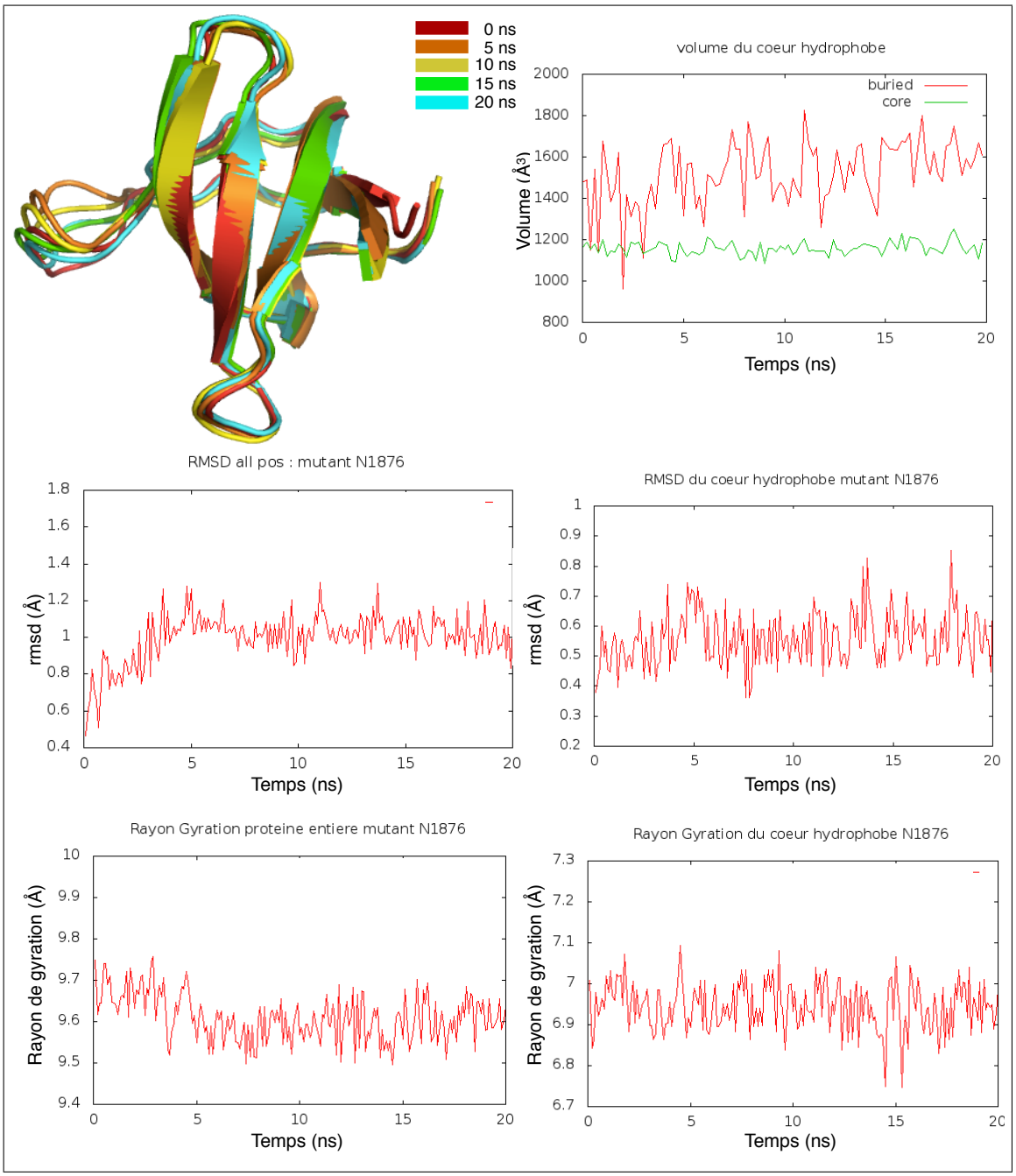


Figure 8.14 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N1876 de la génération Gen2-LIG sans mutation en W_{170} : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.2.3 Protéines mutantes de la génération Gen2-CORE

Enfin, pour les protéines mutantes de la génération Gen2-CORE (Figures 8.15 à 8.19), on peut observer que le RMSD de leur cœur hydrophobe est très proche de celui de la protéine native (en moyenne 0,45 Å). Cela parait assez logique, étant donné que les principales positions du cœur hydrophobe ont été conservées comme natives. Néanmoins, cela signifie également que les autres positions mutées n'exercent pas de contraintes suffisantes pour déstabiliser la protéine. Les RMSD globaux des protéines 34, 428 et 1889 sont assez bons (entre 0,8 et 0,9 Å). Contrairement à la protéine 245 dont la structure des boucles semblent beaucoup plus dévier (RMSD de 1,2 Å).

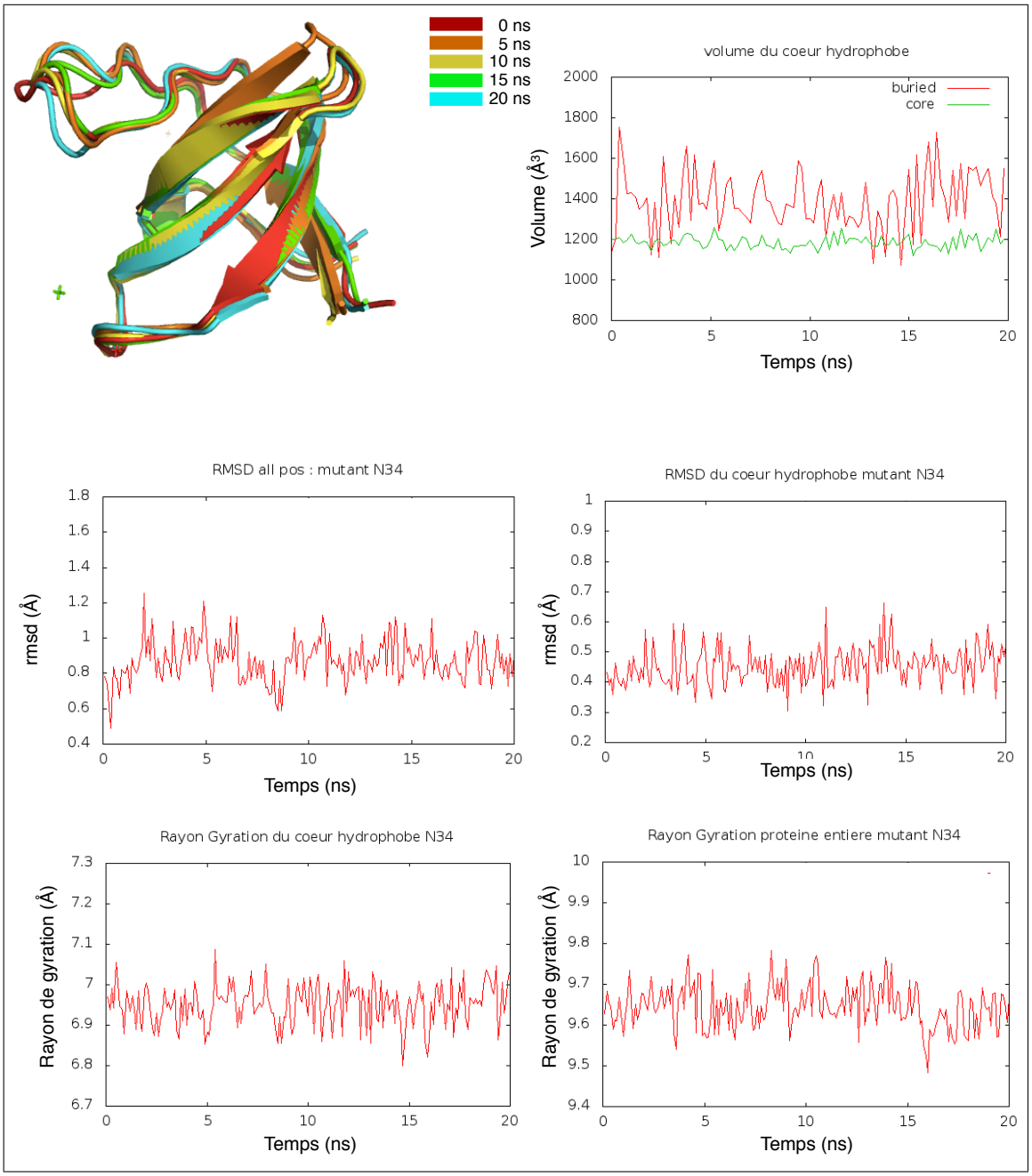


Figure 8.15 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N34 de la génération Gen2-CORE : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

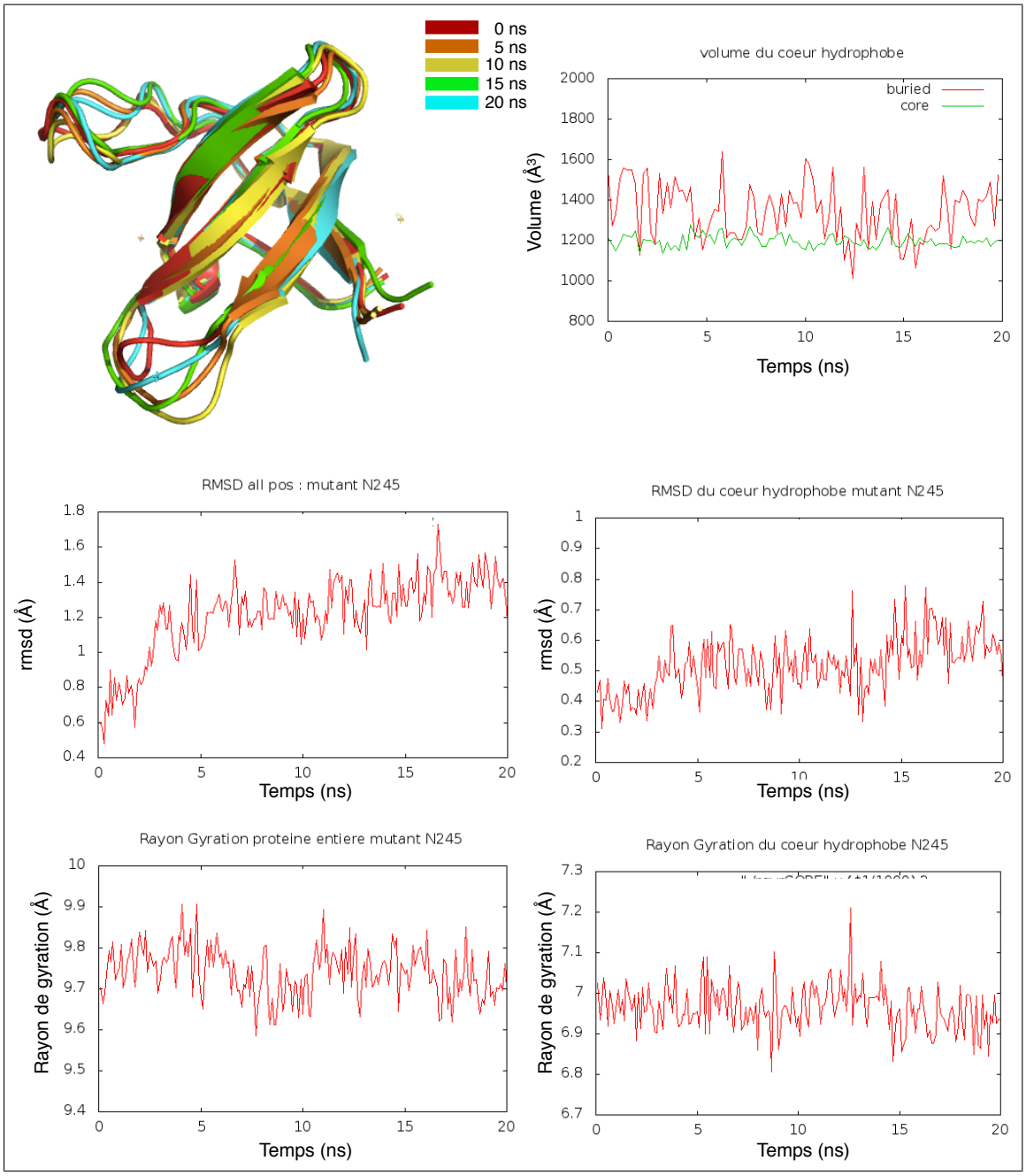


Figure 8.16 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N245 de la génération Gen2-CORE : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

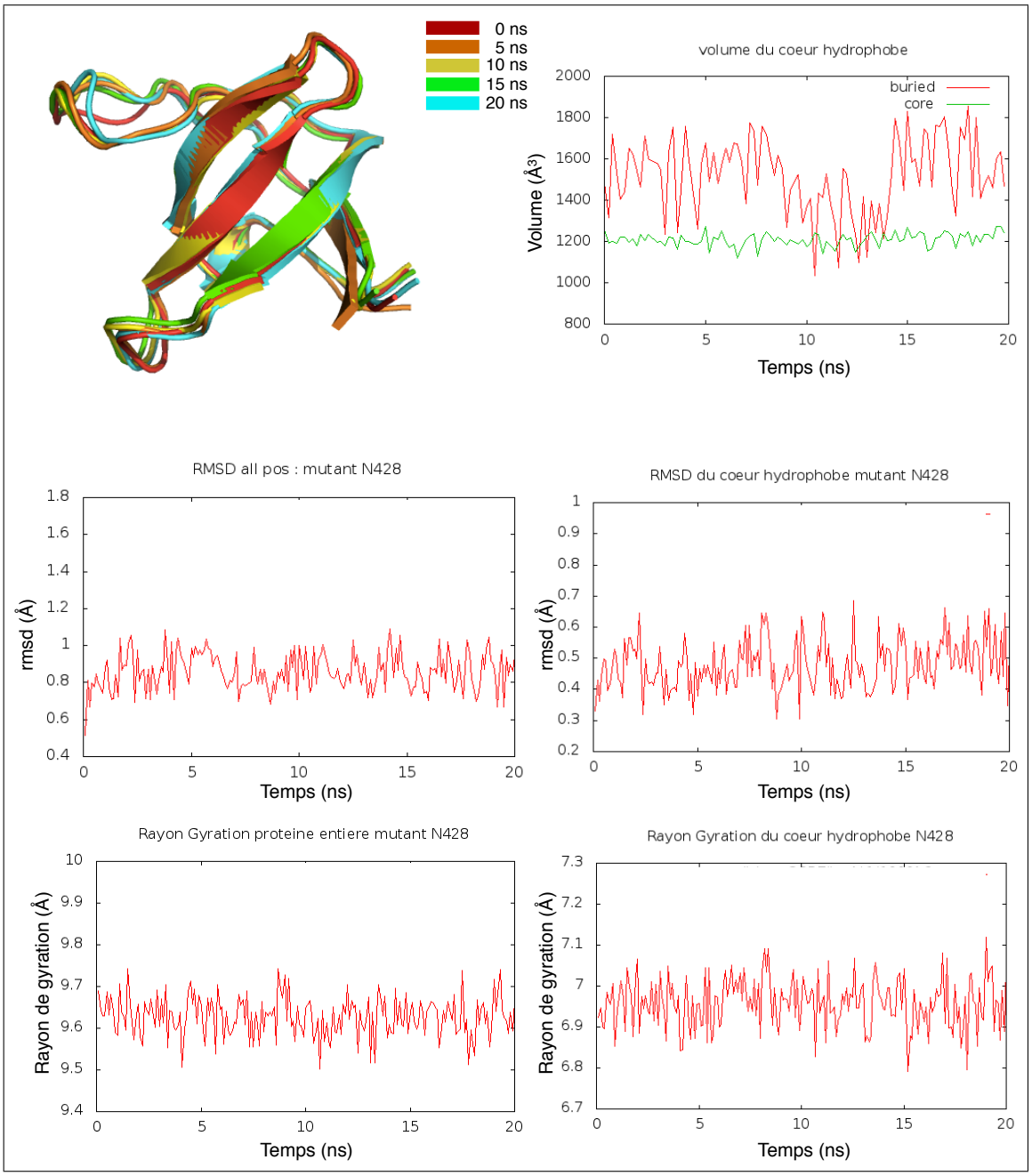


Figure 8.17 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N428 de la génération Gen2-CORE : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

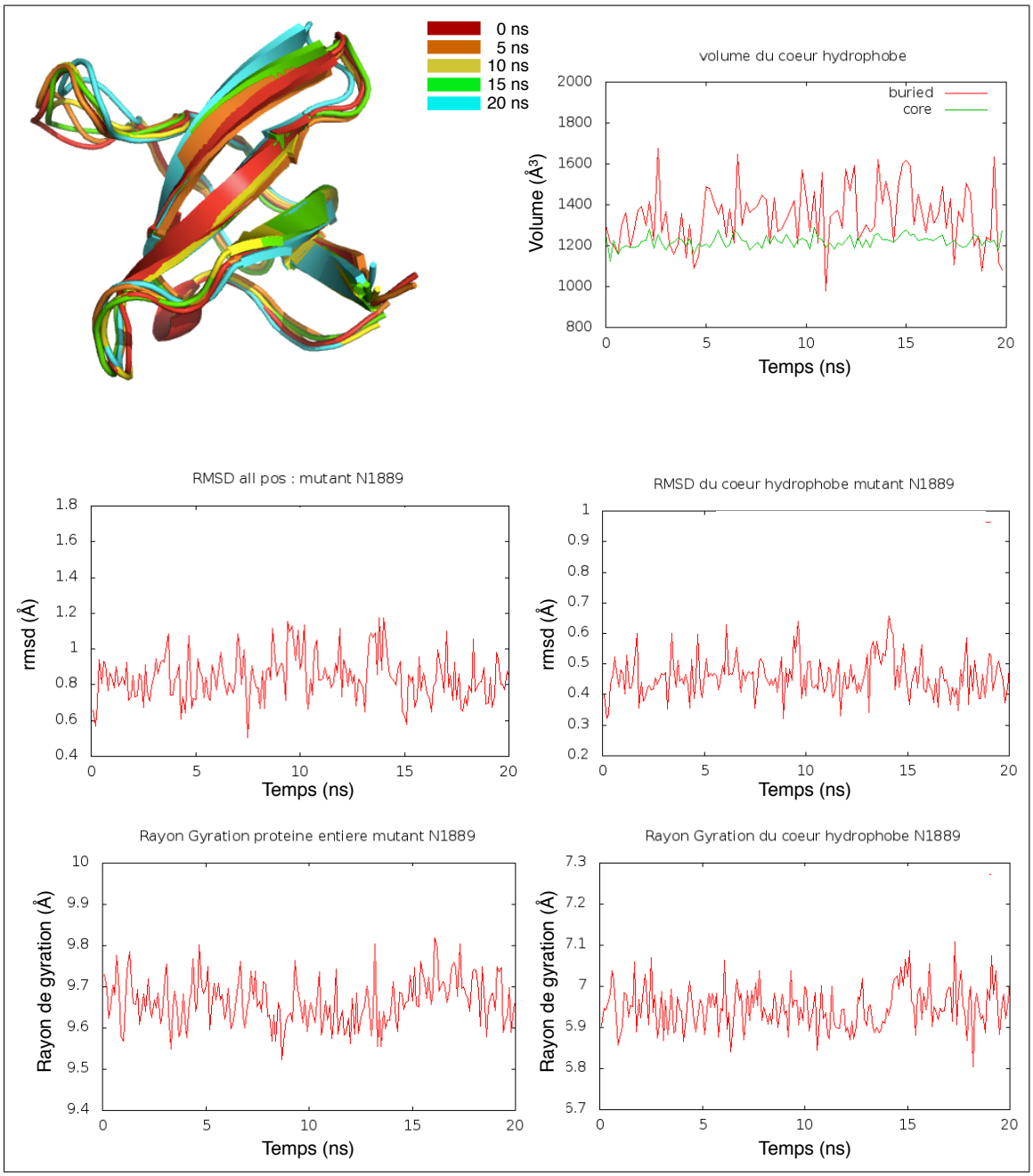


Figure 8.18 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N1889 de la génération Gen2-CORE : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

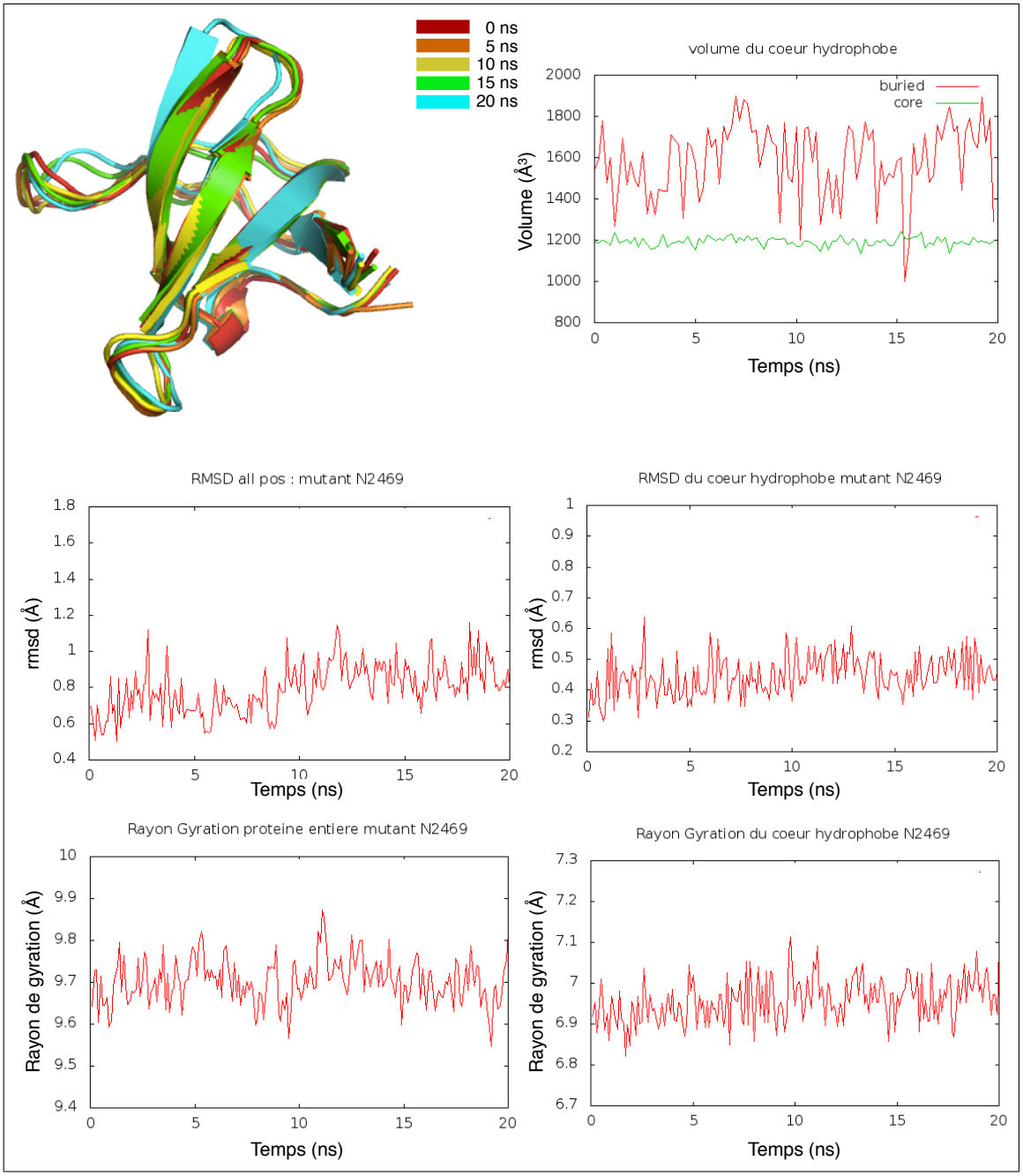


Figure 8.19 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante N2469 de la génération Gen2-CORE : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.2.4 Protéines tests de la génération Gen2-LIG

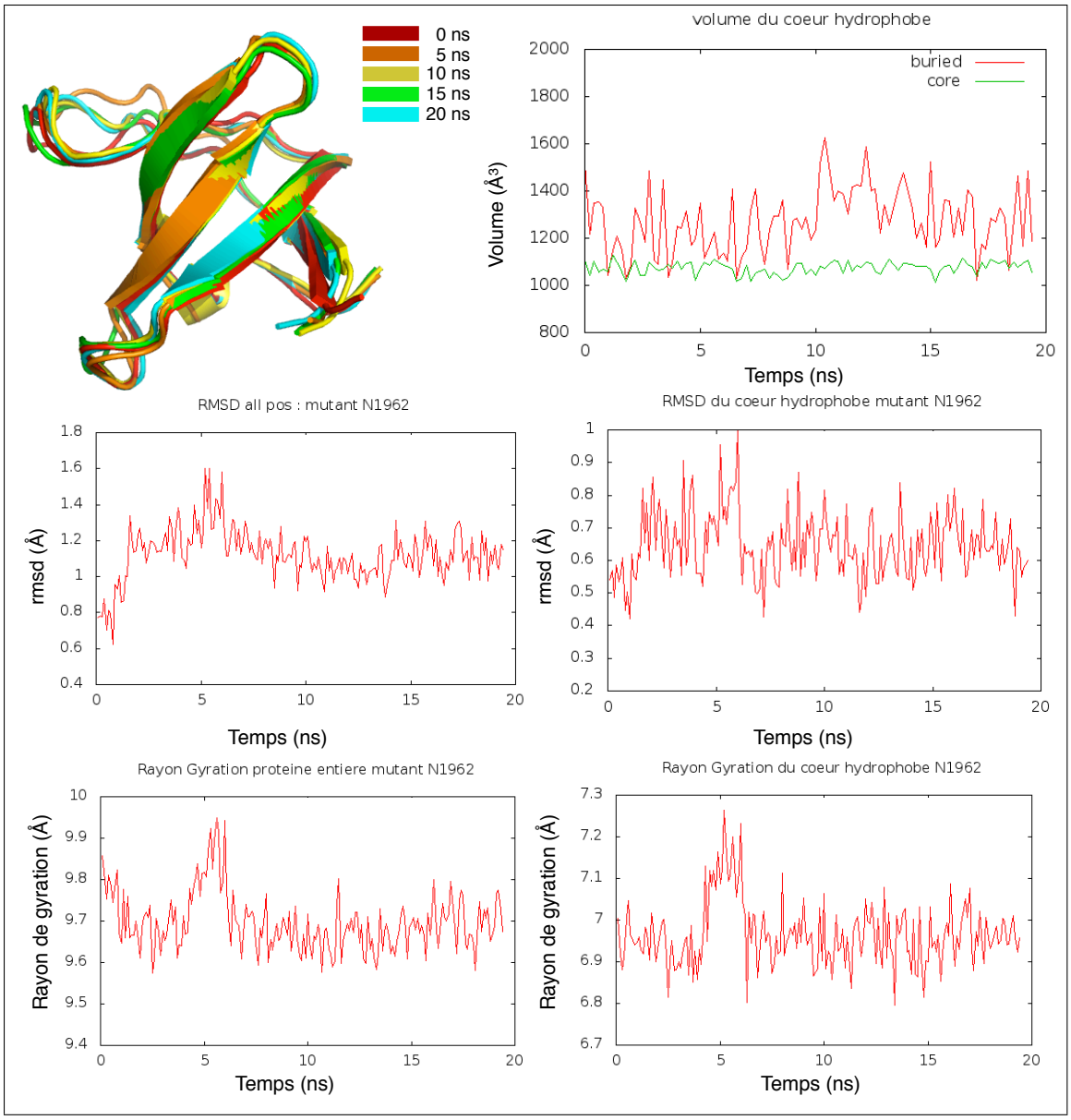


Figure 8.20 – Analyses de la simulation de dynamique moléculaire sur la protéine tests N1962 de la génération Gen2-LIG : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

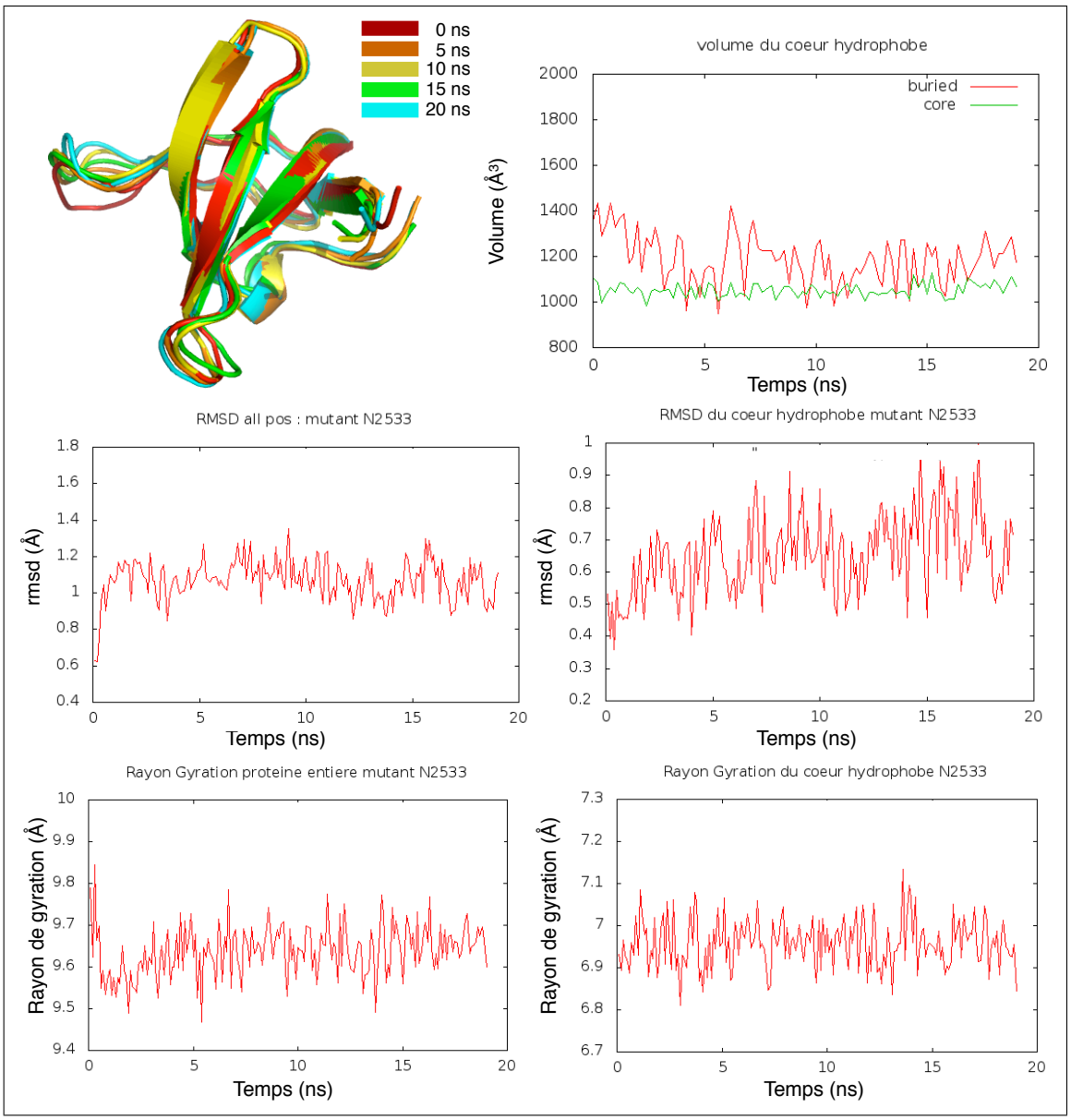


Figure 8.21 – Analyses de la simulation de dynamique moléculaire sur la protéine tests N2533 de la génération Gen2-LIG : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

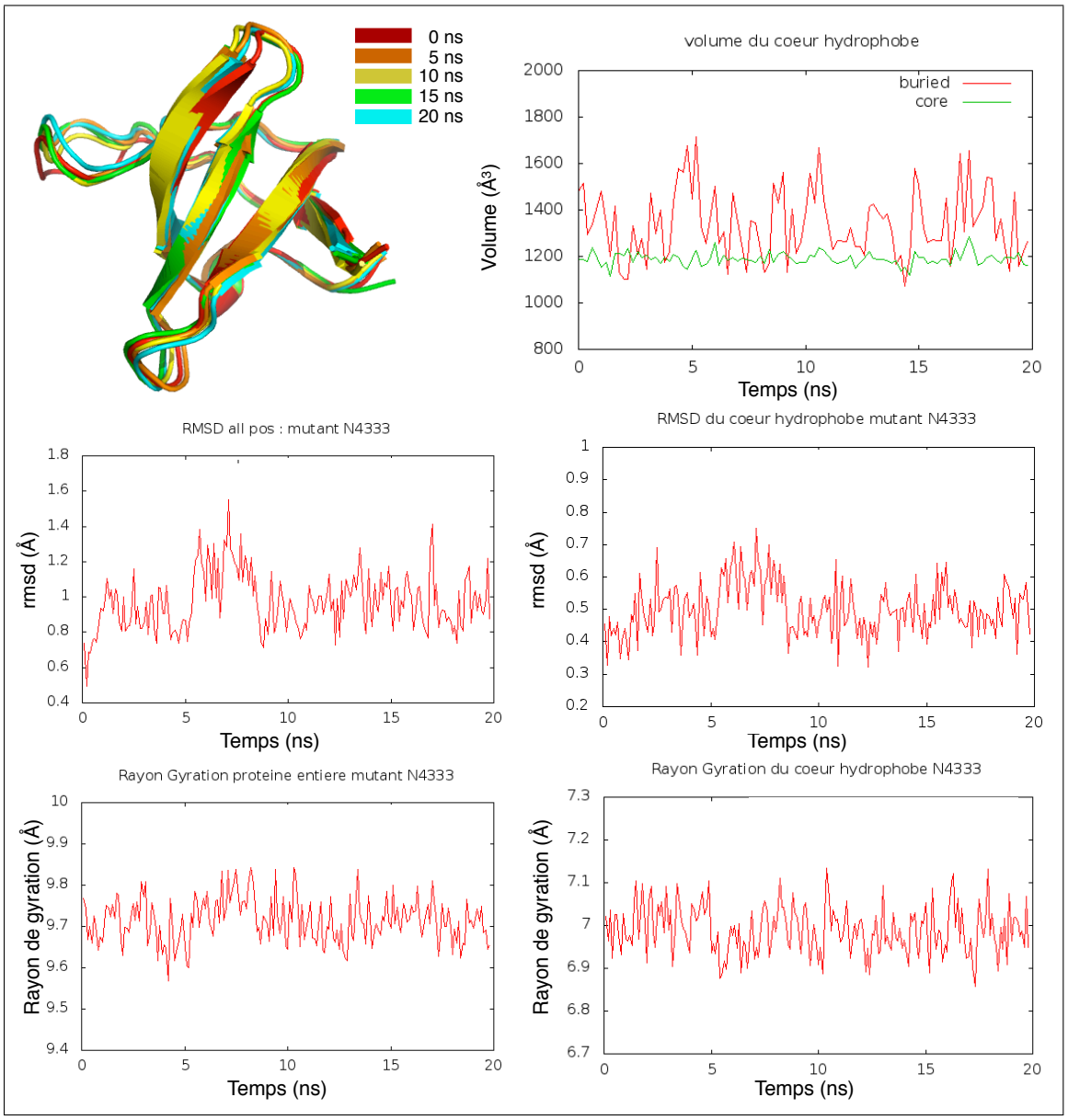


Figure 8.22 – Analyses de la simulation de dynamique moléculaire sur la protéine tests N4333 de la génération Gen2-LIG : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.2.5 Protéine mutante *old* de la génération Gen1

Nous allons comparé ici les comportements des dynamiques des protéines mutantes de la génération Gen2 avec la simulation de dynamique de la protéine mutante *old* de la génération Gen1. La simulation de dynamique moléculaire sur cette protéine mutante a été réalisée exactement dans les mêmes conditions que les simulations réalisées sur la génération Gen2.

D'un point de vue génération, l'ensemble des séquences théoriques sont de qualité équivalente aux générations Gen2-LIG et Gen2-CORE. La différence ici réside dans le choix final des "meilleures" séquences théoriques. En effet, lorsque l'on compare la simulation de dynamique moléculaire (figure 8.23), le comportement de la protéine mutante Gen1-*old* est très éloigné du comportement en simulation de la protéine sauvage, ainsi que des comportements des protéines mutantes précédentes. Le RMSD global atteint presque les 3 Å, c'est-à-dire presque 6 fois plus que pour la protéine sauvage et les protéines Gen2-CORE, et 3 fois plus que les protéines Gen2-LIG. Le RMSD du cœur hydrophobe n'est pas meilleur : il tourne autour des 0,5 Å alors que pour les protéines sauvage et mutantes, c'est globalement 3 fois moins. D'ailleurs, le volume structural du cœur hydrophobe est plutôt de 1300 $Å^3$ au lieu de 1200 $Å^3$ pour les protéines sauvage et mutantes. Les rayons de giration sont très fluctuants ; ils varient entre 9,8 et 10,6 Å pour la protéine entière, et entre 7,3 et 8,2 Å pour le cœur hydrophobe. Ce qui montre que la protéine mutante 1CKA-*old* est bien moins stable que la protéine sauvage, mais également que les autres protéines mutantes générées au cours de cette thèse. La sélection des meilleures séquences théoriques sur la base des descripteurs semble donc assez efficace.

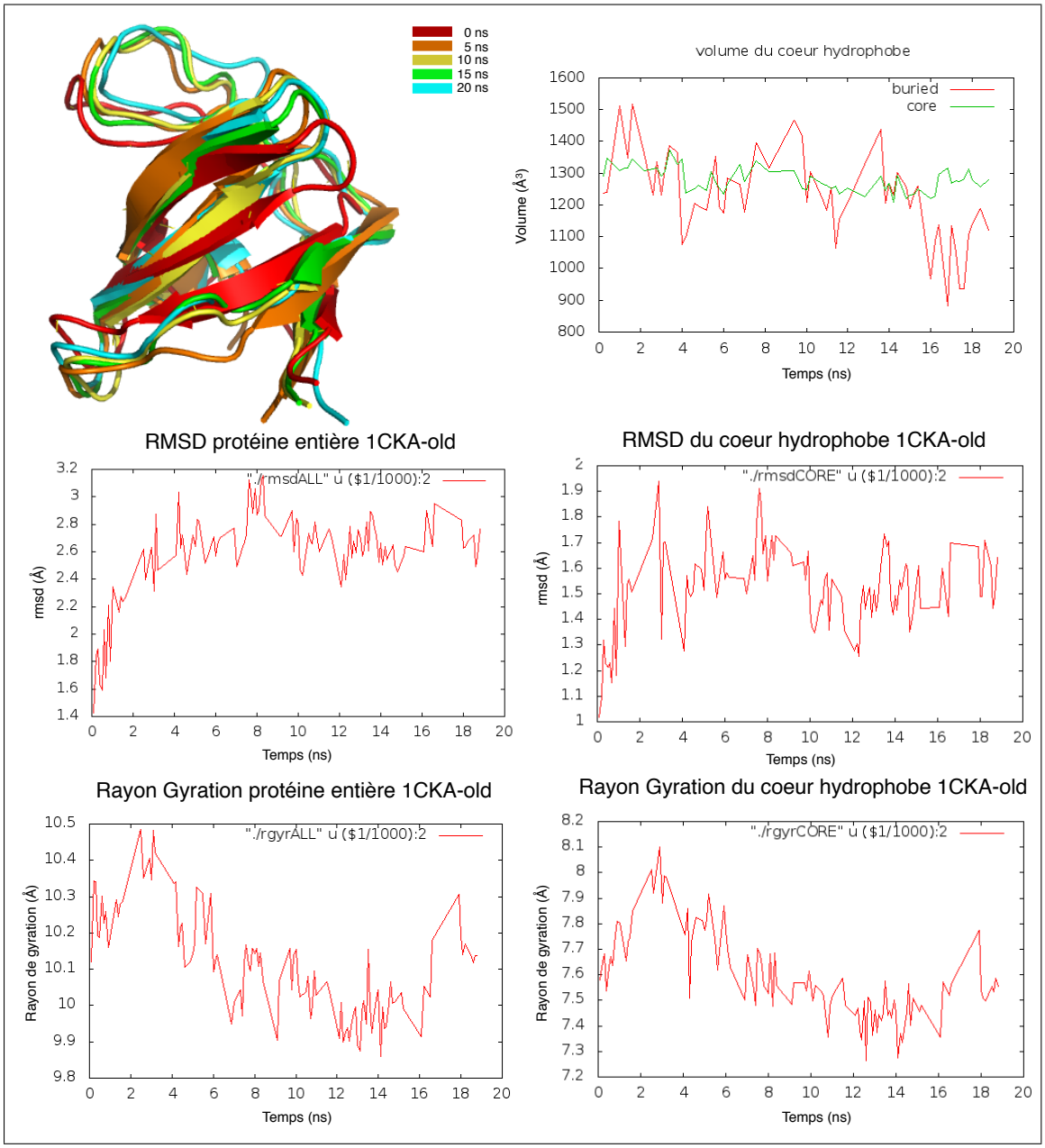


Figure 8.23 – Analyses de la simulation de dynamique moléculaire sur la protéine mutante Gen1-*old* : évolution en fonction du temps (sur 20ns de simulation) du volume du cœur hydrophobe, du RMSD sur toutes les positions et sur les positions du cœur, et du rayon de gyration de la protéine entière et du cœur.

8.3 Résumé des analyses des simulations

Protéine	RMSD All	RMSD Core	Rgyr All	Rgyr Core
WT	0,65	0,4	9,75	6,95
Gen1*old*	2,6	1,5	10,1	7,6
Gen2-LIG-V10	0,9	0,5	9,7	7
Gen2-LIG-V99	0,9	0,6	9,7	7,05
Gen2-LIG-V114	1,1	0,55	9,7	7
Gen2-LIG-V135	1	0,6	9,65	7
Gen2-LIG-V433	1	0,6	9,65	7
Gen2-LIG-V1445	1	0,65	9,75	6,95
Gen2-LIG-V2857	0,9	0,5↗	9,65	7
Gen2-LIG-W105	0,9	0,5	9,6	6,95
Gen2-LIG-W421	0,8	0,5	9,7	6,95
Gen2-LIG-W466	0,8	0,45	9,7	7
Gen2-LIG-W589	1	0,6	9,7	6,95
Gen2-LIG-W847	0,8↗	0,5↗	9,7	6,95
Gen2-LIG-W1876	1	0,55	9,6	6,95
Gen2-LIG-1962	1,1	0,6	9,65	6,95
Gen2-LIG-2533	1	0,8↗	9,6	6,95
Gen2-LIG-4333	0,85	0,45	9,7	6,95
Gen2-CORE-34	0,9	0,45	6,95	9,65
Gen2-CORE-245	1,3↗	0,55↗	9,75	6,95
Gen2-CORE-428	0,8	0,5	9,6	6,95
Gen2-CORE-1889	0,8	0,45	9,65	6,95
Gen2-CORE-2469	0,8	0,45	9,7	6,95

Table 8.1 – Résumé des analyses de simulations de dynamique moléculaire.On y trouve la moyenne et le comportement globale des valeurs des RMSD et des rayons de gyration observés.

8.4 Choix final des mutants pour l'expérimentation

Dans un premier temps, nous allons entreprendre des démarches d'étude expérimentale sur une dizaine de protéines mutantes qui semblent les plus stables en terme de dynamique moléculaire et les plus proches du comportement de la dynamique moléculaire de la protéine sauvage.

Nous avons sélectionné les protéines mutantes 10, 99 et 433 de la génération Gen2-LIG avec la mutation V_{170}. Les protéines 105, 466, 847 et 1876 de la génération Gen2-LIG sans mutation en W_{170}. Et enfin, les protéines 34 et 1889 de la génération Gen2-CORE.

Analyse avec ProSA Nous avons complété ces analyses de nos protéines mutantes avec des outils *online*. Nous avons ainsi évalué la qualité des repliement avec le programme ProSA [Wiederstein & Sippl 2007]. Ce programme compare la structure (fichier PDB) à sa base de données internes de structures 3D et calcule un potentiel énergétique pour chaque résidu : l'énergie de paire (interaction entre paire de carbone C_β), l'énergie de surface (surface des C_β), l'énergie combinée (paire et surface). Dans la figure B.15 (en Annexe), les énergie élevées révèlent des zone de la séquence mal repliée. Dans la figure B.16 (en Annexe), plus les Z-scores sont bas, plus la structure est bonne. Le Z-score dépend de la taille de la protéine. Dans notre cas, si le Z-score de la protéine native 1CKA-WT et les Z-scores des protéines mutantes sont proches, alors d'après ce programme, les structures sont proches. Néanmoins, ici, cette analyse n'apporte pas beaucoup d'information inédites. En effet, nos protéines mutantes ne sont pas sensées changer de conformation globale. Les Z-scores font en fait redondance avec les analyses des RMSD des dynamiques. On peut voir dans la figure B.16 que les Z-scores sont très proches de celui de la protéine sauvage, donc que leur structure sont proches. La méthode ProSA est tout de même séquence-dépendante, aussi, les résultats témoignent d'une répartition assez bonne des résidus polaires et non polaires dans nos séquences prédites.

Analyse avec PSIPRED Nous avons également soumis les séquences sauvage et mutantes à des prédictions de structures : PSIPRED [Jones 1999]. Ici aussi, les résultats semblent prédire un repliement proche de celui de la protéine sauvage.

Analyses biophysiques D'autres outils biophysiques ont été utilisé pour prédire le comportement *in vivo*. Aggrescan [Conchillo-Sole *et al.* 2007] et Tango [Rousseau *et al.*

2006 ; Fernandez-Escamilla *et al.* 2004 ; Linding *et al.* 2004] donnent des probabilités d'agrégation des protéines. Le site ProtParam donne quand à lui des informations sur le point isoélectrique, la stabilité de la protéine, un index aliphatique, et l'hydropathie. Les résultats sur les protéines mutantes des générations Gen1 et Gen2 sont listés dans le tableau en annexe B.1. Nous pouvons comparer ces données avec celles obtenues sur les domaines SH3 sauvages 1CKA, 1CSK, 1UTI, 1SEM et 1ABO (dans le tableau B.2 en annexe). Pour l'index *Aggrescan*, toutes nos séquences théoriques sont prédites comme très sujettes à l'agrégation par rapport aux domaines SH3 sauvages. Seules quelques séquences paraissent meilleures du point de vue de ce critère : Gen1-*old*, Gen2-LIG-V99, LIG-V433, CORE-34, CORE-1889, LIG-1962, LIG-2533 et LIG-4333. Avec le critère *Tango*, les domaines SH3 sauvages sont compris entre 74 et 138, et seules les valeurs de dix séquences sont supérieures, dont Gen2-LIG-1962, 2533 et 4333. Pour l'index de stabilité, il est à prendre avec précaution ; en effet, pour trois sur cinq des domaines SH3 sauvages, il les prédit comme instable. L'index aliphatique semble directement lié avec la composition en résidus GAVLIPM. Et l'index d'hydropathie, les séquences Gen1-*old*, Gen2-LIG-V99, LIG-V433, CORE-34, CORE-1889, LIG-1962 et LIG-2533 semblent plus proches des comportements des protéines sauvages.

8.5 Conclusion

Les simulations de dynamique moléculaire ont montré dans le cadre de cette thèse qu'une sélection poussée et sophistiquée des meilleures séquences grâce à des descripteurs semble être efficace et prometteuse. Pendant les 20 nanosecondes de simulation, les structures restent très proches ($<$1Å) de la structure initiale : la chaîne principale se confond avec celle de la structure native. De plus, la stabilité des protéines mutantes étudiée par dynamique ici, semble être bien meilleure que la stabilité de la protéine mutante 1CKA-*old*.

Chapitre 9

Étude expérimentale sur le domaine SH3-1CKA et plusieurs séquences théoriques

Nous avons tout d'abord étudié expérimentalement la protéine sauvage 1CKA et la protéine mutante 1CKA-*old*. Dans un deuxième temps, la sélection des séquences mutantes dont les comportements bioinformatique et en simulation de dynamique moléculaire qui semblent être des indices encourageants, nous ont permis d'entreprendre des expériences sur de nouvelles protéines mutantes. Le travail expérimental a été fait en collaboration avec Pierre Plateau du Laboratoire de Biochimie de l'Ecole Polytechnique.

9.1 Méthodes d'études structurales

Après une phase de test de surexpression, puis une étape de purification, nous avons envisagé plusieurs méthodes d'étude structurale. Dans un premier temps, la spectroscopie par dichroïsme circulaire nous permettait d'étudier la structure secondaire (hélices, feuillets, etc.) de nos protéines sauvage et mutantes en solution. Cette technique n'utilisant que peu de matériel, et de manière non destructive, la méthode paraissait être idéale dans notre cas.

En parallèle, nous avons mené des expériences de calorimétrie différentielle à balayage (DSC pour *Differential Scanning Calorimetry*) afin d'obtenir des indices thermodynamiques sur la dénaturation de nos protéines : enthalpie calorimétrique de dénaturation, enthalpie effective de dénaturation et température de dénaturation T_d.

Grâce à l'optimisation d'un protocole de production fournissant des quantités plus importantes de protéines purifiées (sauvage et mutante), nous avons pu envisager des méthodes de détermination de structures plus poussées. Les deux techniques privilégiées pour l'étude structurale de protéines à une résolution atomique sont la cristallographie aux rayons X et la spectroscopie de résonance magnétique nucléaire (RMN).

La première technique requiert la formation d'un monocristal de protéine sur lequel est projeté un faisceau intense de rayons X. L'analyse du motif de diffraction résultant permet de reconstituer la densité électronique de la molécule et de disposer les hétéroatomes dans l'espace. Néanmoins, cette méthode peut présenter des inconvénients. L'analyse du signal n'est pas immédiate car l'expérimentateur n'accède à la phase du signal que de manière indirecte (substitution isomorphe multiple d'atomes lourds, diffraction anomale multiple, remplacement moléculaire à partir d'une structure analogue). Par ailleurs, les régions désordonnées de la molécule ne contribuent pas au motif de diffraction et n'apparaissent pas dans la structure finale. Et surtout, l'obtention d'un monocristal constitue l'une des étapes limitantes de cette technique.

La seconde technique est basée sur l'analyse des signaux de résonance magnétique émis par les noyaux d'atomes individuels de la protéine après excitation par une impulsion de radiofréquences. Une grande variété d'expériences permet d'obtenir des informations sur la structure (connectivité, distances inter-protons, angles dièdres, orientations dans un repère lié à la molécule) et sur la dynamique de la protéine (sur une gamme d'échelles de temps allant de 10 à 103 s). Une des étapes limitantes de la RMN est l'attribution de chaque signal de résonance au(x) noyau(x) d'atome correspondant(s), étape préalable à toute interprétation précise des spectres. Par ailleurs, l'analyse des spectres est rendue

difficile par l'importance des recouvrements de signaux sur une largeur spectrale limitée. Enfin, la relaxation plus rapide du signal pour les protéines de grande taille limite la RMN à l'étude des protéines de taille modeste (quelques dizaines de kDa).

La cristallographie par rayons X a été testée dans un premier temps sur protéine mutante 1CKA-*old*. Néanmoins, c'est une méthode longue, fastidieuse et souvent vaine. Ce qui nous a encouragé à nous focaliser sur une méthode de résolution structurale plus rapide et plus facile à mettre en place : la spectroscopie de résonance magnétique nucléaire (RMN). Cette méthode nous est apparue comme adéquate pour notre cas d'étude. En effet, de part sa petite taille (6-7 kDa), la protéine 1CKA est une protéine modèle pour l'étude structurale par RMN.

9.1.1 Étude par dichroïsme circulaire

Le dichroïsme circulaire (CD) [Kelly *et al.* 1751] s'appuie sur la capacité des molécules qui ont une activité optique : propriété d'absorber différemment la lumière polarisée. Cette technique non destructive va nous permettre d'obtenir de premiers indices sur l'état de structuration de nos protéines mutantes, comme la proportion de chaque type de structures secondaires dans nos protéines dans différentes conditions (températures), et de comparer la forme générale des spectres CD entre les protéines sauvage et mutantes.

Une solution contenant un soluté qui absorbe la lumière absorbe celle-ci selon la loi de Beer-Lambert :

$$A = log\left(\frac{I}{I_0}\right) = \varepsilon cl \tag{9.1}$$

où A est l'absorbance du soluté (ou densité optique), I_0 l'intensité de la lumière incidente à une longueur d'onde donnée λ, I l'intensité de la lumière transmise à λ, ε le coefficient d'extinction molaire du soluté à λ, c sa concentration molaire et l la longueur

en cm du trajet optique. La valeur de ε varie avec λ ; un graphique de ε en fonction de λ pour le soluté est son spectre d'absorbance.

Les polypeptides absorbent fortement la lumière dans la région ultraviolette (UV) du spectre ($\lambda = 100$ à 400 nm) essentiellement parce que leurs chaînes latérales aromatiques (celles des Phenylalanines, des Tryptophanes et des Tyrosines) ont des coefficients d'extinction molaire particulièrement élevés dans cette région du spectre. Cependant, les polypeptides sont incolores, car ils n'absorbent pas la lumière visible ($\lambda = 400$ à 800 nm).

Pour les molécules chirales comme les protéines, les valeurs de ε sont différentes pour la lumière polarisée en cercle vers la gauche (ε_L) ou vers la droite (ε_R). La variation, avec λ, de la différence de ces valeurs, $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$, donne le spectre CD du soluté étudié. Dans le protéines, les hélices α, les feuillets β, et les enroulements au hasard ont des spectres CD caractéristiques (Figure 9.7). Le spectre CD d'un polypeptide donne donc une idée de sa structure secondaire.

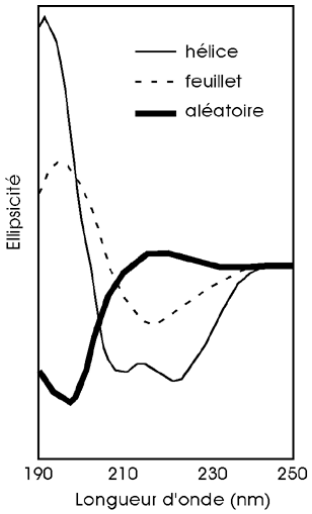


Figure 9.1 – Spectres CD caractéristiques des structures secondaires : hélice α, feuillet β et boucles.

9.1.2 Étude par calorimétrie

Le DSC donne accès, de façon directe, à l'information thermodynamique des protéines dans un intervalle de température donné. Cette technique offre une description thermodynamique complète de l'état natif d'une protéine et des états intermédiaires qui apparaissent lors d'une dénaturation par la température. Ces paramètres thermodynamiques de dénaturation nous permettent de comprendre les mécanismes de stabilisation de nos protéines.

La dénaturation thermique des protéines en solution est un mécanisme au cours duquel certaines liaisons faibles, responsables de la structure native des protéines, sont rompues. Suite à ce mécanisme, l'architecture initiale des protéines est transformée en une structure spatiale plus désordonnée. Les protéines globulaires, ayant une architecture initiale compacte sont concernées par ce mécanisme qui est réversible à faible concentration et en l'absence d'interaction entre les protéines. Dans le cas contraire, le mécanisme de dénaturation s'accompagne d'une phénomène d'agrégation selon le schéma suivant :

$$P_N \underset{k_N}{\overset{k_D}{\rightleftharpoons}} P_D \xrightarrow{k_A} A$$

avec : P_N la concentration en protéine dans un état natif, P_D celle dans un état dénaturé, k_D et k_N des constantes de réaction de dénaturation ou de renaturation, et k_A la constante de réaction d'agrégation. Si $k_A < k_N$, la réaction de dénaturation est considérée comme réversible.

9.1.3 Étude par RMN

9.1.3.1 Méthode générale

La résonance magnétique nucléaire (RMN) permet la détermination de structure tridimensionnelle de macromolécules biologiques, telles des protéines qui sont, préférentiellement, inférieures à 30 kDa. Bien que la RMN est été découverte il y a plus de 60 ans,

l'utilisation de la RMN pour étudier la structure des protéines est récente et se développe plus particulièrement depuis le début des années 1980 [Voet & Voet 1998].

Le principe de la RMN est le suivant : les noyaux atomiques dotés d'un nombre impair de protons, de neutrons ou des deux, auront un spin nucléaire intrinsèque. Cette méthode utilise les propriétés magnétiques de ces noyaux, comme ceux ayant un spin 1/2 tels que le ^{1}H, ^{15}N et ^{13}C. Ces noyaux ont la particularité d'avoir deux états quantiques qui ont la même énergie en absence de champ magnétique. Par contre, en présence d'un champ magnétique, il en résulte une différence des deux états qui est en fonction du noyau et de l'intensité du champ. En effet, la RMN est basée sur l'utilisation de champs magnétiques intenses et des transitions entre les différents états quantiques. De plus, l'environnement local autour d'un noyau donné, dans une molécule, a tendance à perturber légèrement le champ magnétique externe et à affecter son énergie de transition. Cette dépendance de l'énergie de transition vis-à-vis de la position d'un atome particulier dans une molécule rend la RMN extrêmement utile pour la détermination de la structure des molécules [Wolf & Haken 2004].

En présence d'un champ magnétique, les spins des noyaux ^{1}H, ^{15}N et ^{13}C, s'alignent préférentiellement dans la direction du champ et oscillent à une fréquence qui dépendra du champ magnétique ressentit (la fréquence de Larmor). De la somme des spins individuels résulte une aimantation globale qui sera parallèle au champ externe. L'application d'une radio fréquence d'une durée appropriée pour une impulsion de 90 degrés résulte entraîne une aimantation globale qui est perpendiculaire au champ externe. Cette aimantation transverse oscille à la fréquence de Larmor et peut être enregistrée. Le signal enregistré est une FID (*Free Induction Decay*) ; une oscillation correspondant à une fonction sinus qui décroît vers zéro en raison du retour à l'équilibre. Le rapport entre la fréquence de résonance d'un noyau et celle d'une référence correspondra au déplacement chimique (δ) [Evans 1995].

Cette technique, tout comme la diffraction des rayons X, exige une concentration élevée de protéines dans les échantillons, une pureté et une homogénéité exemplaires de ces macromolécules, et elles doivent être stables, du moins, suffisamment pour la durée de l'enregistrement des spectres.

Dans notre étude structurale des domaines SH3 (sauvage et mutantes), les premiers spectres enregistrés seront des spectres 1D-^{1}H ou 2D-^{15}N-HSQC. Les autres spectres RMN multidimensionnels impliquant un ou plusieurs types de noyaux différents ne seront pas envisagés, faute de temps pour mener à bien ses expériences.

9.1.3.2 Spectre 2D ^{15}N-HSQC

La réalisation d'un spectre 2D ^{15}N-HSQC (pour *Heteronuclear Single Quantum Coherence*) Bodenhausen & Ruben [1980], corrèle les protons et les atomes de ^{15}N de chaque groupement amide de la protéine. Son résultat peut être visualisé sous forme de spectres à deux dimensions où le déplacement chimie des protons est porté en abscisse et celui des atomes d'azote en ordonnée. Dans une protéine, chaque résidu est caractérisé par la liaison N–H de sa fonction amide, à l'exception des prolines (dont l'azote ne porte aucun hydrogène) et du résidu N-terminal (l'hydrogène de la fonction amine étant labile). Chacune des liaisons amides N–H d'une protéine enrichie en ^{15}N donne lieu à un pic de corrélation sur le spectre HSQC ^{1}H-^{15}N.

La position du pic dans le spectre est décrite par les déplacements chimiques du proton et de l'azote. Le déplacement chimique traduit le « blindage » d'un noyau par les électrons, s'opposant à l'établissement du champ magnétique. Il dépend de tous les facteurs qui peuvent modifier la structure ou la géométrie des orbitales moléculaires. Plusieurs interactions physiques interviennent dont les contributions ne sont pas encore toutes comprises. L'interprétation du déplacement chimique reste donc essentiellement qualitative. On peut généralement distinguer [Oldfield 1995] :

- Les effets à courte portée : longueurs et angles de liaison (nature chimique du noyau), angles de torsion (d'ou l'interprétation des déplacements chimiques en termes de structure secondaire), liaisons hydrogène fortes...
- Les effets à longue portée : contributions électrostatique (dipôles voisins) et magnétique (courant de cycle aromatique, anisotropie du carbonyle...). Ces interactions sont particulièrement sensibles à la structure tertiaire de la protéine et cette sensibilité est plus forte pour le proton que pour les autres noyaux.

Le spectre HSQC ^{15}N constitue une empreinte de la protéine d'étude et de son état de repliement, d'où son utilisation fréquente dans les études de faisabilité en RMN [Yee et al., 2002].

Les spectres de RMN dépendent des caractéristiques du tampon dans lequel la protéine est diluée (notamment sa force ionique et son pH) et de la température à laquelle est réalisée l'expérience. Il était donc important de choisir dès le début un tampon adapté aux expériences envisagées.

9.2 Résultats

Dans le cadre de cette thèse, une partie expérimentale a été menée afin de valider le modèle de prédiction de séquences théoriques. En effet, l'observation d'une similitude entre la structure d'une protéine mutante prédite et choisie par nos soins et la structure de départ (de la protéine sauvage) *in vitro* serait un argument solide pour valider le modèle bioinformatique. Le schéma 9.2 résume les étapes expérimentales envisagées et réalisées lors de cette thèse.

Après la génération de séquences théoriques à partir de la structure 1CKA du domaine SH3, nous avons sélectionné quelques protéines mutantes d'après nos filtres (décrits dans les chapitres précédents), traduit et optimisé en codon d'ADN afin de commander les gènes synthétisés et intégrés dans des plasmides.

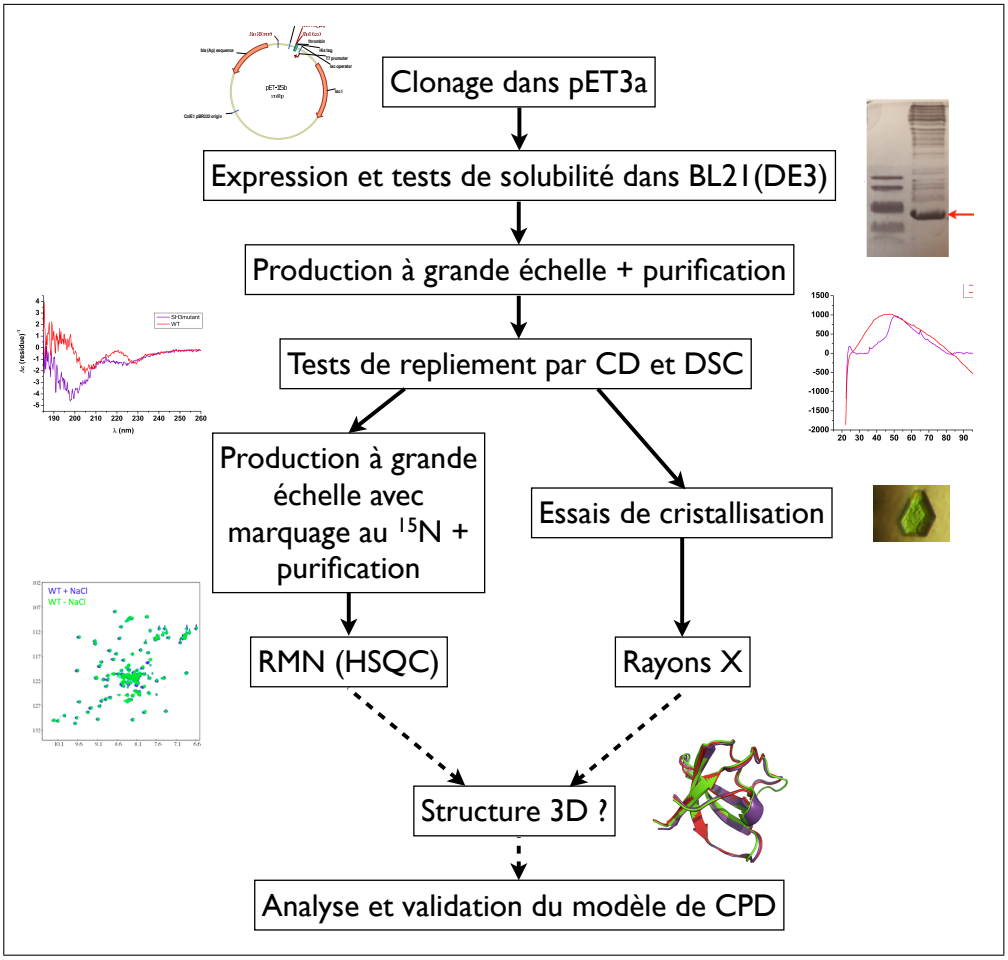


Figure 9.2 – Du gène à la structure protéique : procédure expérimentale dans le projet de validation de notre modèle CPD. Les flèches en pointillés n'ont été que des étapes envisagées et non réalisées. On peut voir la carte du plasmide pET3a, un gel d'électrophorèse montrant la 1CKA-WT, les spectres de dichroïsme circulaire et de calorimétrie de 1CKA-WT et *old*, un spectre HSQC-2D de 1CKA-WT, des cristaux obtenus, la structure 3D de la protéine SH3-WT.

Après une étape de clonage des gènes d'intérêt dans le plasmide pET3a, les plasmides sont ensuite transformés dans une souche d'*E. coli* afin d'exprimer nos protéines. Ensuite le procédé dépend de la solubilité de la protéine mais les étapes principales restent la production à grande échelle puis la purification ainsi qu'une étape de spectrométrie de masse permettant de savoir si on dispose de la bonne protéine. Une fois la quantité satisfaisante et la pureté acceptable obtenues, des expériences de dichroïsme circulaire, de calorimétrie et de RMN sont entrepris pour connaître son état de repliement.

9.2.1 Protéine sauvage 1CKA-WT et protéine mutante 1CKA-*old*

Nous nous sommes appuyé sur les travaux de Wu *et al.* [1995] sur la surexpression de la protéine sauvage 1CKA pour reproduire les conditions et se servir des résultats comme référence et point de comparaison.

9.2.1.1 Clonage

Les gènes codant pour la protéine sauvage et la protéine 1CKA-*old* ont été synthétisé par Epoch Biolabs, dans un plasmide pBluescript II SK(-). Nous les avons clonés dans un plasmide pET15b puis dans un plasmide pET3a entre les sites de restriction NdeI et BamHI (voir les cartes des plasmides en annexe et la figure 9.3). Dans les plasmides pBuescript et pET15b, les gènes étaient précédés d'une étiquette de 6 histidines (His_6Tag) afin de la purifier par affinité. Pour le plasmide pET3a, nous avons fait le choix de retirer le HisTag.

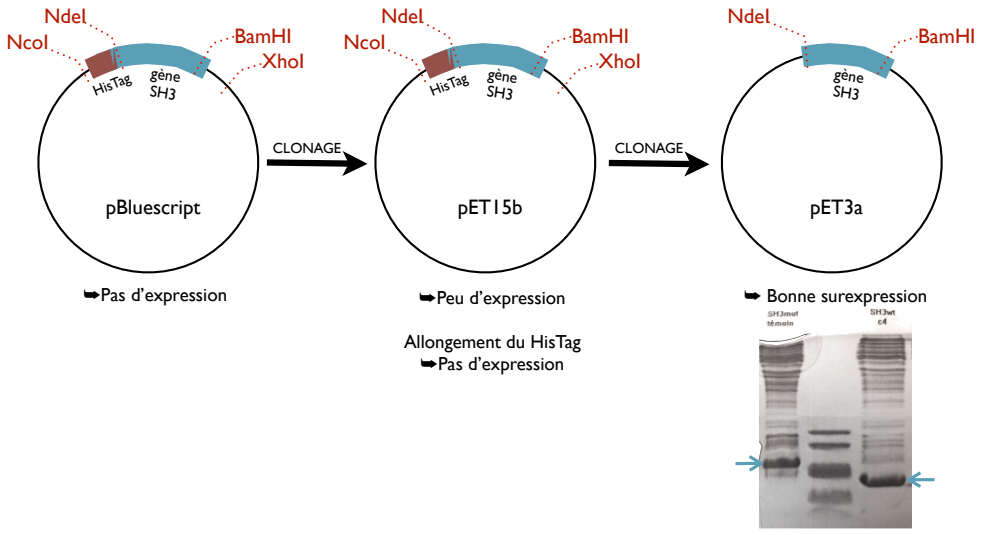


Figure 9.3 – Étapes de clonage et résultats de test de surexpression pour 1CKA-WT et 1CKA-*old*.

Préculture	Culture	Induction	Résultats
5mL LB + 100mM Amp	*ensemenc. à 0,05DO/mL dans 1L LB + Amp*	*à 0,6DO/mL + 50mM IPTG*	
37 ° C ON	37 ° C	37 ° C 3h	Bonne expression
37 ° C ON	37 ° C	30 ° C 3h	Peu d'expression
37 ° C ON	37 ° C	20 ° C 3h	Pas d'expression
37 ° C ON	30 ° C	30 ° C ON	Pas d'expression
-	37 ° C	37 ° C ON	Pas d'expression
-	induction directe	37 ° C ON	Pas d'expression

Table 9.1 – Liste des conditions de culture testées pour la production de la protéine sauvage 1CKA et la protéine mutante Gen1-*old*. *(ON : sur la nuit)*

9.2.1.2 Production et purification

Dans le cas du plasmide pBluescript, nous n'avons réussi à produire aucune des deux protéines. Avec le gène dans le plasmide pET15b, nous obtenions une production assez faible. Aussi pour améliorer la purification sur talon (colonne de chromatographie à affinité Cobalt) nous avons rallongé le HisTag par PCR à 10 histidines. Nous en avons conclu d'après la réalisation de tests de surexpression vains, qu'un HisTag trop long semblait perturber le repliement du domaine. Nous avons donc décidé de supprimer l'étiquette HisTag, et de cloner le gène dans le plasmide pET3a.

Nous avons réussi à optimiser les conditions de surexpression pour la protéine sauvage SH3-1CKA et pour la protéine mutante 1CKA-*old* dans la souche BL21(DE3) (voir le détail du protocole en annexe), comme nous pouvons le voir dans les gels d'électrophorèse des figures 9.4 et 9.5. Nous listons les différentes conditions de culture testées dans le tableau 9.1.

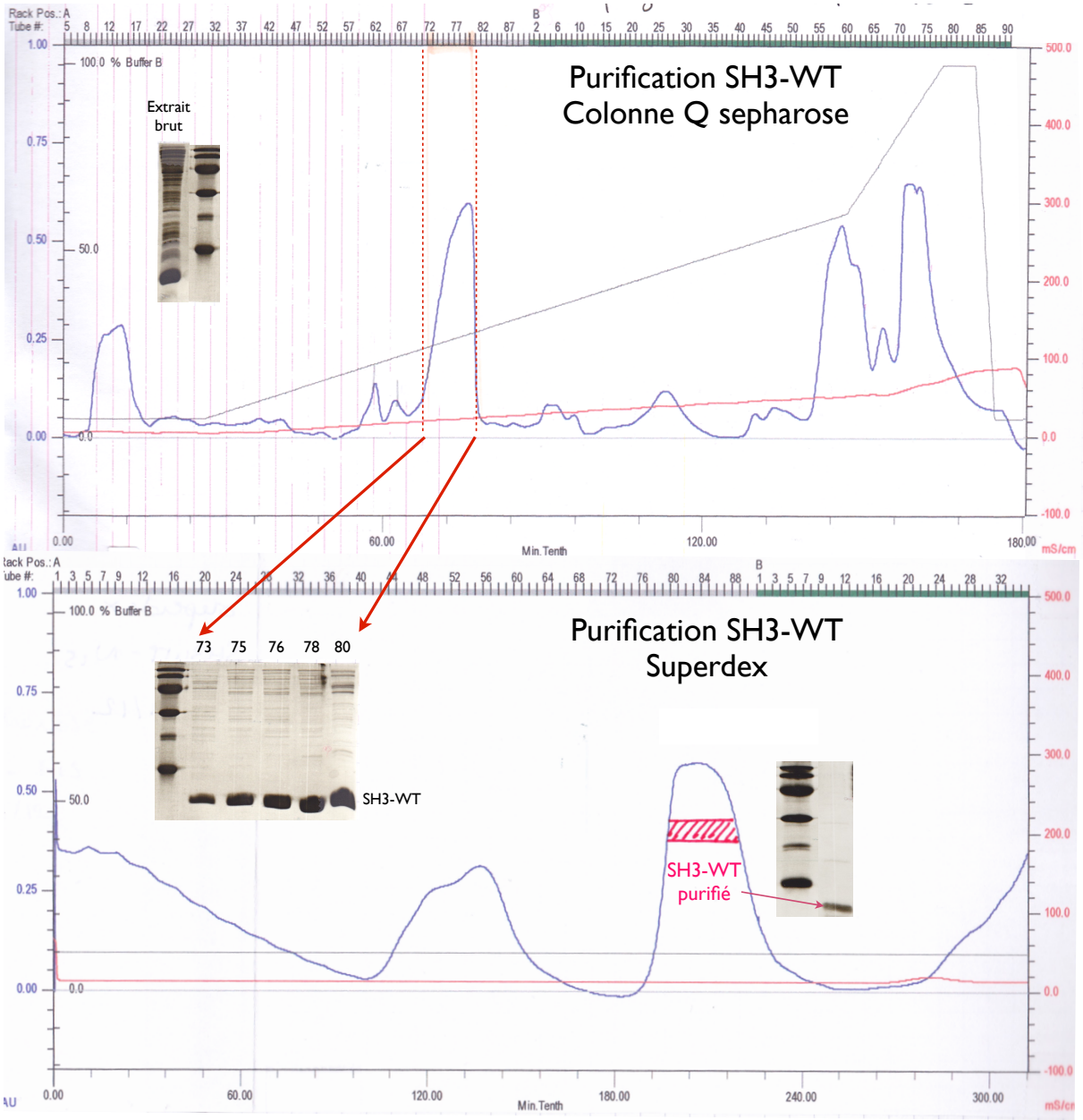


Figure 9.4 – Étapes de purification et gels SDS-PAGE pour la protéine 1CKA-WT.

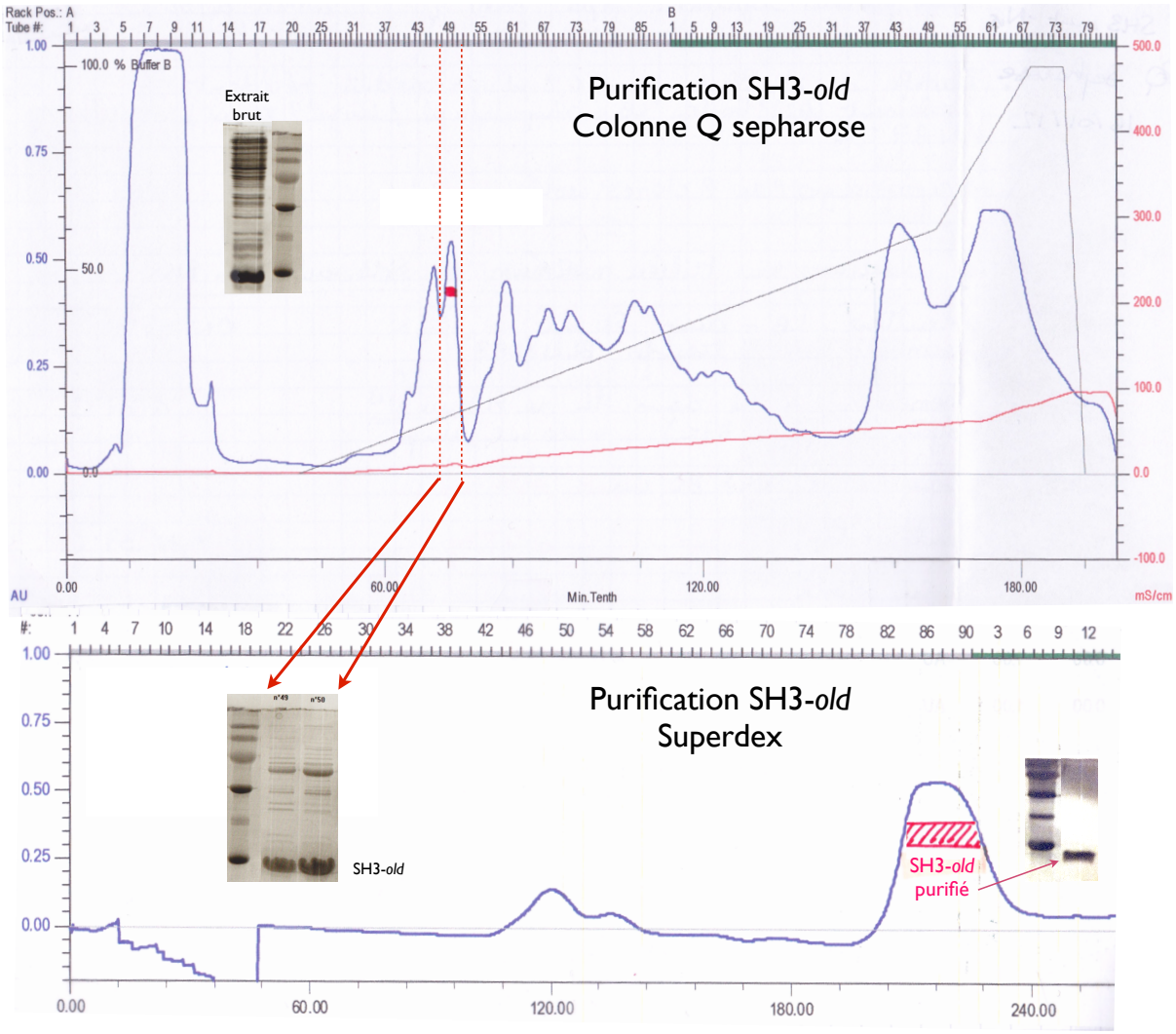


Figure 9.5 – Étapes de purification et gels SDS-PAGE pour la protéine 1CKA-*old*.

9.2.1.3 Dichroïsme circulaire et calorimétrie

Ayant une quantité suffisante de protéine sauvage purifiée, nous avons pu réaliser des expériences de calorimétrie et de dichroïsme circulaire (CD) (spectropolarimètre Jasco) avec l'aide de M. Valerio-Lepiniec du laboratoire IBBMC d'Orsay. Les spectres de dichroïsme circulaire sont compris entre 185 nm et 260 nm, et les échantillons sont préparés avec des tampons 20 mM NaH_2PO_4. Nous disposions de 22 μM de protéine sauvage, et 17 μM de protéine mutante 1CKA-*old*. Pour les expériences de calorimétrie (DSC), nous disposions 0,5 mg/ml de 1CKA-WT, et 0,3 mg/ml de 1CKA-old dans un tampon 10 mM Hépès pH 7,5 et 0,1 mM EDTA.

Résultats typiques des domaines SH3 en CD Dans les travaux de Tanaka *et al.* [2011] et Angrand *et al.* [2001], les spectres obtenus en dichroïsme circulaire sur des domaines SH3 (codes PDB : 1QWF et 1SHG) sont très caractéristiques des protéines à feuillets β [Opatowsky *et al.* 2003], comme nous pouvons le voir dans la figure 9.6 : un minimum à 200 nm et un maximum à 220 nm. En effet, le minimum à 217 nm est typique des β-protéines, mais le maximum à 220 nm semble dépendre de l'environnement des résidus aromatiques de la protéine.

Résultats de dichroïsme circulaire Comme nous pouvons le constater dans les figures 9.7, la protéine sauvage 1CKA-WT présente à 20 °C des structures secondaires typiques des domaines SH3 : deux minimums à 207 nm et 230 nm, et un maximum vers 190 nm. Lorsque la protéine sauvage est chauffé à 90 °C, le spectre de dichroïsme change de forme : conséquence de la dénaturation des structures secondaires. En revenant à la température de départ, 20 °C, le spectre retrouve sa forme initiale, indiquant donc que le repliement des structures secondaires est réversible.

Dans le cas de la protéine mutante 1CKA-*old*, son spectre à 20 °C est différent de celui de la protéine sauvage et semble nettement moins structuré. L'amplitude du spectre

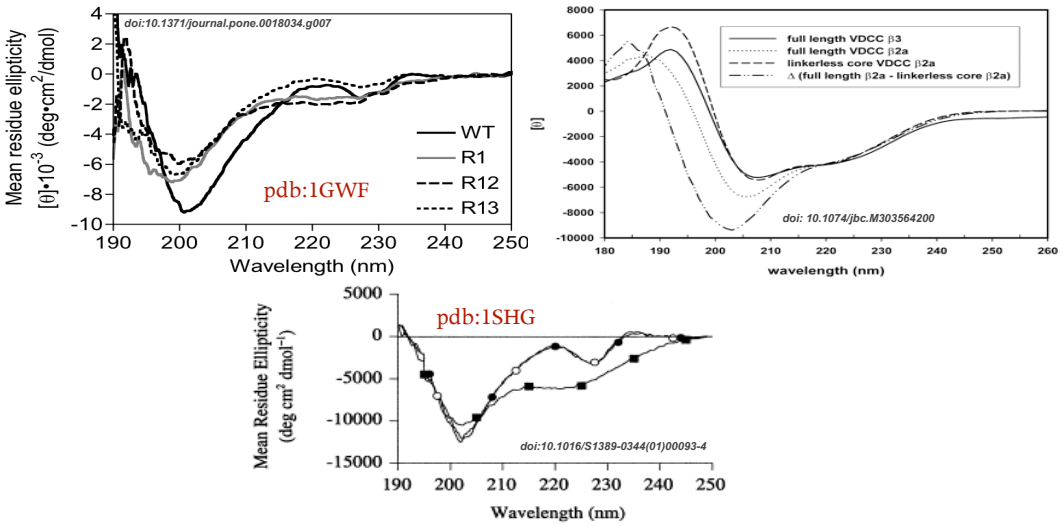


Figure 9.6 – Spectres de dichroïsme circulaire : *à droite*, typiques des domaines SH3 ; *à gauche*, caractéristiques de protéines en feuillet β.

s'atténue lorsque l'on chauffe à 90 ° C, et il reprend sa forme initiale en revenant à 20 ° C. Ce qui signifie tout de même qu'il n'y aurait pas un niveau de structuration complètement nulle.

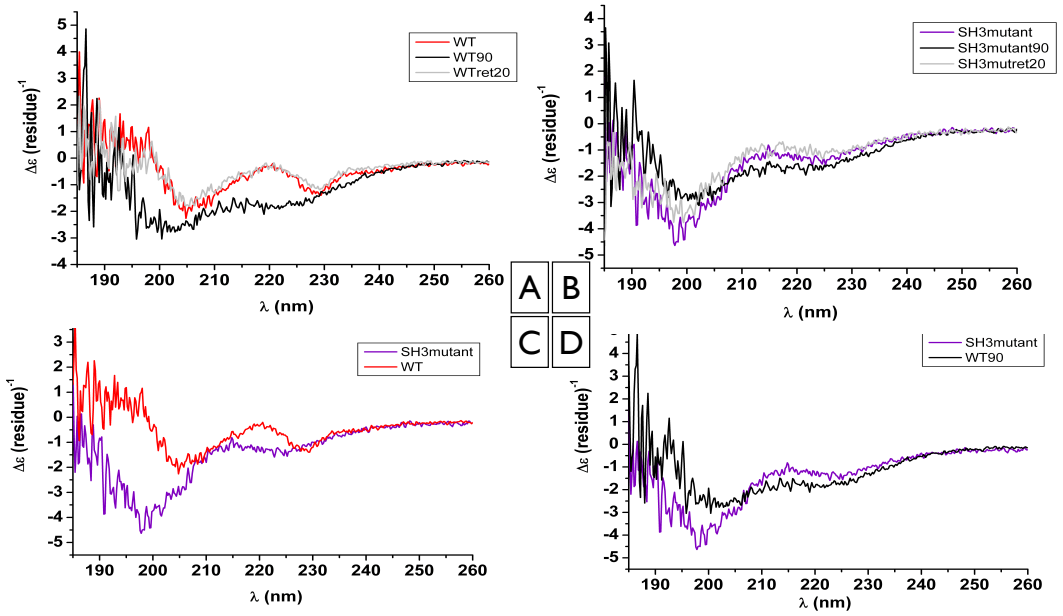


Figure 9.7 – Spectres de dichroïsme circulaire sur la protéine sauvage 1CKA-WT et la protéine mutante 1CKA-*old*. **A :** Spectres avec 1CKA-WT à 20 ° C (non dénaturé), puis dénaturé à 90 ° C, et renaturé à 20 ° C. **B :** Idem pour 1CKA-*old*. **C :** Comparaison des spectres à 20 ° C pour 1CKA-WT et 1CKA-*old*. **D :** Comparaison des spectres de 1CKA-WT dénaturé à 90 ° C et 1CKA-old non dénaturé.

Résultats typiques des domaines SH3 en DSC D'après de nombreuses études sur les domaines SH3, dont les travaux de Filimonov *et al.* [1999], Morel *et al.* [2006] et Casares *et al.* [2004], montrent des spectres de calorimétrie assez similaires, comme nous pouvons le voir sur la figure 9.8.

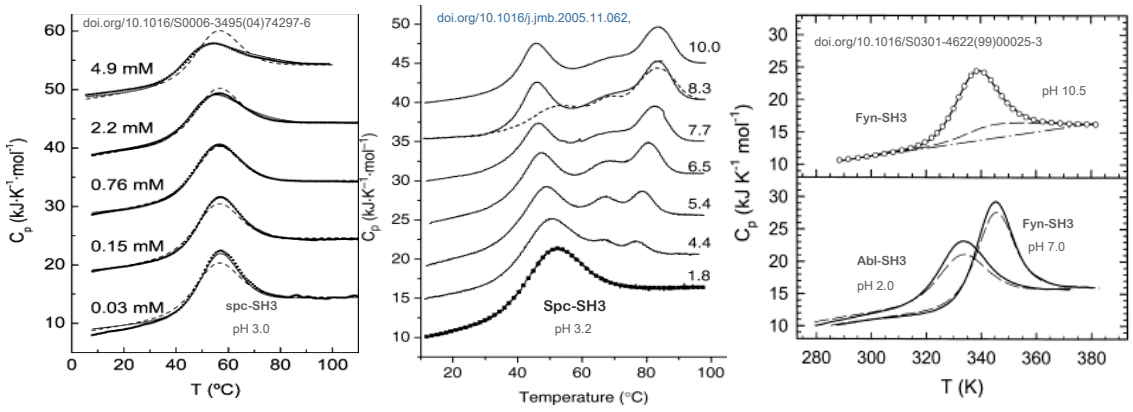


Figure 9.8 – Spectres de calorimétrie pour des domaines SH3 de différentes protéines et à des pH différents.

Résultat de calorimétrie Les thermogrammes de la figure 9.9 montrent que très peu de choses. Nous pouvons supposer soit que les protéines sont peu structurées soit qu'il n'y avait pas assez de quantité de ces protéines de petites tailles pour obtenir un signal correct. Néanmoins, nous pouvons observer une amplitude de 1 Kcal aux alentours de 50 ° C, ce qui parait cohérent par rapport aux valeurs typiques des domaines SH3.

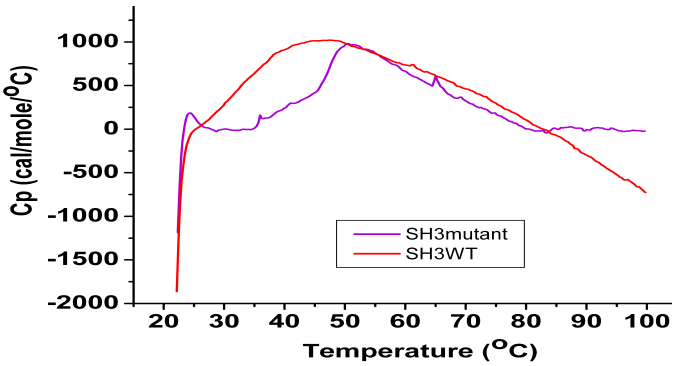


Figure 9.9 – Spectres de calorimétrie pour 1CKA-WT et 1CKA-*old*.

9.2.1.4 Spectre RMN-HSQC 2D

Résultats typiques des domaines SH3 en RMN Voici quelques spectres RMN-2D typiques de domaines SH3, d'après les travaux de Vaynberg & Qin [2006] et Kang *et al.* [2000a] (figure 9.10).

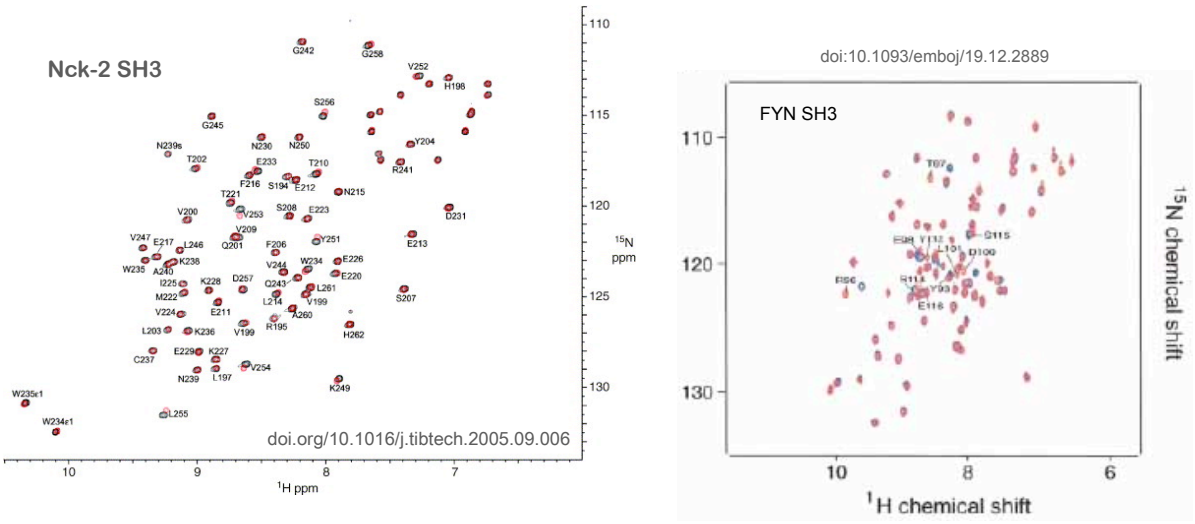


Figure 9.10 – Spectres RMN HSQC 2D pour des domaines SH3.

Résultat de RMN Nous avons pu fournir une grande quantité de protéine sauvage purifiée marquée au ^{15}N afin de réaliser des expériences de RMN avec notre collaborateur I. Guijarro de l'Institut Pasteur.

Nous pouvons voir dans la figure 9.11 que le spectre de la protéine sauvage montre deux espèces plus ou moins équimolaires : l'une structurée et l'autre désordonnée. De plus, elles sont semblables avec et sans sel. Étant donné que le nombre de signaux est environ le double de celui attendu (97 pics au lieu de 50), l'hypothèse la plus probable est un mélange à l'équilibre de protéine structurée et désordonnée du à un échange lent chimique.

Le spectre de la protéine mutante 1CKA-*old* sans sel correspond au spectre d'une protéine pas ou peu structurée (environ 45 signaux au lieu de 50 attendus). Les signaux

des groupements amides montrent une faible dispersion, des échanges rapides avec l'eau, donc une exposition au solvant (spectres a-b-c de la figure 9.11).

Malheureusement, l'échantillon de la protéine mutante avec sel a été contaminé par la protéine sauvage (spectre d dans la figure 9.11).

La superposition du spectre à pH 6 avec 0,2 mM de Na2SO4 et celui à pH 7.5 sans sels de la protéine mutante (spectres e et f dans la figure 9.11), montrent que ces spectres sont très similaires et présentent une faible dispersion des déplacements chimiques. Quelques signaux se sont déplacés, ainsi que quelques uns "nouveaux", que l'on ne voyait pas à pH 7.5 probablement dû à un échange trop rapide avec les protons de l'eau. Les expériences hetsofast indiquent qu'à pH 6, la protéine mutante n'est pas structurée. Les données de relaxation du noyau 15N sont aussi en accord avec une protéine très dynamique, non globulaire. Enfin, les expériences noesy et roesy montrent très peu de signaux comme attendu pour une protéine désordonnée, et, on n'observe aucun signal qui puisse indiquer l'existence d'un feuillet beta pour la protéine mutante 1CKA-*old*.

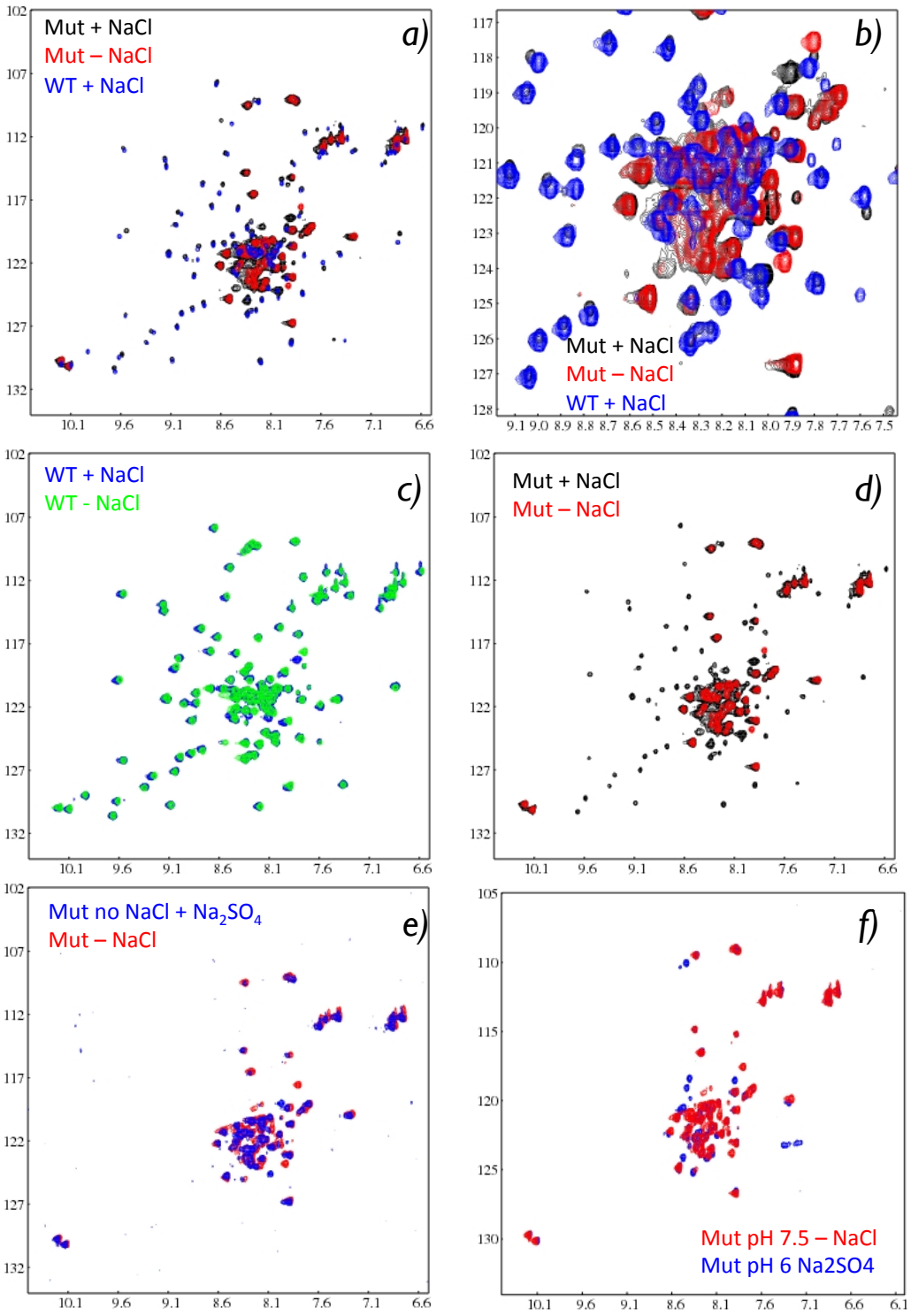


Figure 9.11 – Spectres RMN HSQC 2D pour 1CKA-WT et *old*. **a) :** spectres pour la protéine sauvage en présence de sel, de la protéine mutante avec et sans sel. **b) :** zoom sur les spectres de a). **c) :** comparaison des spectres pour 1CKA-T en présence ou non de sel. **d) :** idem que c) pour la protéine mutante. **e) et f) :** spectres pour la protéine mutante sans sel en changeant le pH.

9.2.2 Protéines mutantes 1CKA-LIG et 1CKA-CORE

Les gènes codant pour ces protéines mutantes 1CKA-LIG et 1CKA-CORE, ainsi que pour la protéine sauvage 1CKA-WT ont été synthétisés par MWG Eurofins, dans un plasmide PCR2.1 TOPO. Nous avons réalisé des tests de surexpression sur ces plasmides. Nous avons réussi à surexprimer seulement la protéine mutante LIG-W847, production confirmée par spectroscopie de masse. Néanmoins, la protéines se trouvait dans la partie non soluble. De plus, nous avons rencontré des difficultés à reproduire cette production nous empêchant de faire d'autres expériences. Nous avons voulu cloner dans le plasmide pET3a entre les sites de restriction NdeI et BamHI. Mais là aussi, nous avons rencontré des difficultés.

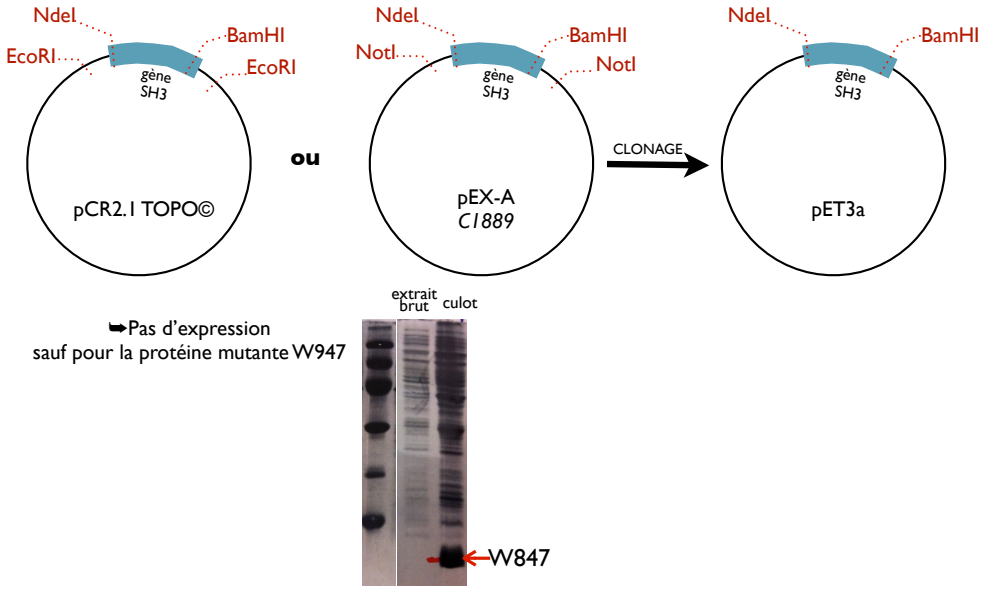


Figure 9.12 – Résumé des étapes surexpression et de clonage pour les protéines mutantes 1CKA-LIG et 1CKA-CORE.

9.2.3 Autres protéines mutantes

Nous avions prédit des séquences théoriques à partir d'un domaine SH2 de structure 1BM2. Nous avions utilisé les mêmes paramètres de génération que ceux pour Gen1. De

cette génération, nous avons sélectionné une protéine mutante 1BM2-*old*. Nous n'avons pas réussi à surexprimer cette protéine, suggérant qu'elle était peu structurée.

9.3 Conclusion sur l'étude expérimentale des séquences théoriques

Les premiers mois de cette thèse, m'ont permis d'apprendre et de maîtriser les méthodes expérimentales nécessaires. Nous avons pu surexprimer et purifier la protéine sauvage 1CKA-WT, afin d'avoir un point de comparaison lors d'expériences de détermination structurale. En parallèle, nous avons trouvé les conditions optimales de surexpression et de purification de la protéine mutante Gen1-*old*. Plusieurs indices ont laissé penser que 1CKA-*old* était bien structurée : sa bonne surexpression et solubilité, puisqu'elle échappe à la protéolyse et à l'aggrégation, ainsi que les expériences de dichroïsme circulaire et calorimétrie. Malheureusement, le criblage d'un bon nombre de conditions de cristallogénèse sur cette protéine mutante n'a rien donné. Le marquage au N15 nous a par contre permis de caractériser par RMN (spectre 2D) l'état de structuration de la protéine 1CKA-*old*. Il nous ait apparu clairement que la protéine sauvage était structurée et que la protéine mutant était peu structurée, ou en *molten globule* (état stable de protéine en partie repliée) [Pande & Rokhsar 1998]. L'hypothèse qu'elle soit complètement déstructurée est réfutée par les résultats de dichroïsme circulaire, où le changement de température implique une déstructuration puis un retour à l'état d'origine.

L'équipe de Renaud Vincetelli (AFMB) à Marseille continue nos expériences sur les protéines mutantes de la génération Gen2. Néanmoins, les constructions choisies intégrant les gènes d'intérêt dans des plasmides commerciaux semblent apporter une toxicité.

Le travail bioinformatique nous a fourni des séquences mutantes en principe plus prometteuses sur le plan expérimental. Néanmoins, seule une protéine mutante n'a pu être

surexprimée, et nous avons eu du mal à reproduire les conditions. Cependant, cela reste encourageant pour la suite des recherches.

Chapitre 10

Conclusion et discussions

Le CPD cherche à identifier ou prédire des séquences d'acides aminés compatibles avec un repliement donné de la chaine polypeptidique. Le CPD a des applications majeures pour l'ingénierie des protéines, mais aussi pour la prédiction de structures. Grace au CPD nous pouvons prédire un ensemble de séquences prédites à partir d'une structure donnée. Pour ajuster les paramètres et les protocoles de génération de séquences, nous avons exploré deux approches différentes. Cette thèse a donc porté sur deux aspects principaux. Le premier était l'application du CPD pour la prédiction de structures. Nous avons donc mis au point une analyse originale pour rechercher des homologues naturels dans les bases de données génomiques. Le deuxième aspect était la validation des outils et méthodes qui sont développées dans notre équipe pour le CPD. Cette validation comprend une composante bioinformatique : caractérisation des séquences prédites. Elle comprend aussi une composante expérimentale : production et caractérisation de quelques une de ces séquences prédites, pour tester leur bon repliement.

La méthode d'évolution dirigée pour générer les séquences théoriques semble assez fiable. Cependant, il existe quelques limitations. Cette méthodes considère les boucles comme des structures fixes et bien définies. Les coordonnées spatiales du squelette peptidique ne bougent que très peu pendant l'algorithme. Pourtant *in vivo*, les boucles sont justement très libres, sans perturber la stabilité de la molécule et contraignant peu ou pas la nature de leurs acides aminés. Ainsi, la génération de séquences produit des interactions

biaisées entre résidus situés dans les boucles. Il faudrait peut-être envisager dans l'avenir d'intégrer à notre algorithme d'évolution dirigée une méthode modifiant dans une certaine limite, les coordonnées spatiales du *backbone* au niveau des boucles (ou plus).

L'analyse des covariances nous a permis d'identifier les interactions entre acides aminés. La localisation 3D des ensembles de positions révélés par cette analyse semble valider notre méthode d'un point de vue biologique. En effet, pour plusieurs structures d'un même domaine, on a pu constater que les ensembles d'acides aminés trouvés par l'analyse spectrale de chacune se superposent. Les positions des ensembles sont généralement très proches structurellement, formant des réseaux de résidus qui stabiliseraient le rapprochement de plusieurs structures secondaires. Cependant, le choix des résidus reste encore très heuristique et demanderait peut-être le soutien d'une méthode encore plus rigoureuse, voire mathématique. Mais nous ne pouvons pas oublier l'aspect biologique qui nous oblige à garder une phase manuelle pour sélectionner ces positions.

La classification des acides aminés semble simplement gommer les covariances entre acides aminés spécifiques et fait ressortir plus particulièrement les covariances entre types d'acides aminés dues à leurs propriétés physico-chimiques

Grâce à l'analyse des covariances, les recherches Blast montrent que ces groupes améliorent toujours le nombres d'homologues trouvés. Néanmoins, les résultats de recherche d'homologues sur les séquences théoriques sont bien moins bons que sur des séquences naturelles, de l'ordre de moins 80%. Les séquences prédites par CPD sont assez éloignées en terme d'identité des séquences naturelles (seulement 30% d'identité). Malgré ces quelques limitations, ces résultats suggèrent que notre méthode pourrait être utilisée pour identifier de nouveaux membres de ces familles de protéines.

Une partie de ces résultats ont été décrite dans un article publié en 2010 (Schmidt am Busch, Sedano, Simonson, Plos One 2010) ; l'ensemble devrait faire l'objet d'un article en 2013 (Sedano, Mignon, Schmidt am Busch, Steyaert, Simonson, en préparation).

Dans la deuxième partie de la thèse, nous avons effectué de nouvelles générations de séquences théoriques par CPD pour produire de nouvelles variantes de la protéine SH3-1CKA-wt. En effet, les premières générations de séquences dans lesquelles nous avons sélectionné les protéines mutantes Gen1-*old* et SH2-1BM2-*old* avaient été faites avec un modèles physique assez simple et une première version de notre logiciel CPD.

Nous avons trouvé les conditions expérimentales optimales pour surexprimer et purifier la protéine sauvage 1CKA-wt. En parallèle, nous avons réussi à reproduire l'une des protéines prédites, Gen1-*old*. Plusieurs indices ont laissé penser que cette protéine mutante était bien structurée : sa bonne surexpression et solubilité montrant qu'elle échappait à la protéolyse et à l'agrégation, et des expériences de dichroïsme circulaire et de calorimétrie. La caractérisation structural de Gen1-*old* par cristallographie n'a rien donné mais son marquage à l'azote N15 et sa caractérisation par RMN a révélé que son état de structuration était quasi nulle. Nous avons également cloné et tenté d'exprimer 1BM2-*old*. Ces tentatives ont été infructueuses, suggérant qu'elle aussi était peu structurée.

Les nouvelles générations de séquences en utilisant un modèle physique amélioré, et une paramétrisation plus récente et soignée, nous ont permis de réaliser une analyse bioinformatique afin de mesurer la qualité des séquences prédites par notre méthode et pour identifier des candidats pour de nouvelles expériences. La mise en place d'une série de descripteurs nous a permis de filtrer et de sélectionner 19 séquences candidates qui, sur le papier, ont des caractéristiques nettement améliorées par rapport aux premières protéines mutantes testées Gen1-*old* et SH2-1BM2.

Les simulations de dynamique moléculaire sur des durées de 20 nanosecondes sur ces protéines mutantes ont révélé que leurs structures restaient très proches (inférieur à 1 Å) de la structure initiale. Cependant, 20 nanosecondes de simulation est trop court pour avoir un témoin crédible de stabilité. Ici, le but était d'avoir un premier indice sur la stabilité et le comportement des protéines mutantes. Dans l'avenir, on pourra allonger le

temps de simulation pour être proche du dépliement réel du domaine SH3, de l'ordre de 1000 ou 2000 nanosecondes.

Dans la phase suivante, nous avons tenté d'exprimer la moitié de ces protéines candidates, et seule la protéine mutante W847 a été produite, malheureusement en trop petite quantité pour caractériser sa structure. Elle semblait de toutes façons peu soluble, témoignant peut-être d'une structuration moindre. Les approches expérimentales n'ont pas abouti par suite de diverses difficultés techniques qui ne paraissent pas insurmontables, mais qui restaient relativement longues à résoudre dans le cadre de cette thèse. Mais ces expériences seront continuées sur les protéines mutantes de la génération Gen2. L'équipe de R. Vincetelli se concentre dessus mais ils n'ont réussi jusqu'à présent qu'à caractériser la protéine sauvage 1CKA. Ces expériences sont importantes pour valider le modèle CPD de génération de séquences. Une fois qu'un certain nombre de structures de protéines mutantes auront été caractérisées comme identiques

Ces protocoles d'analyse et de validation semblent être de bons moyens pour caractériser nos séquences théoriques générées par CPD. En effet, l'analyse des covariances permet dans un premier temps d'évaluer la qualité des séquences prédites, et dans un deuxième temps d'améliorer la recherche d'homologues et à terme d'en faire un outil pour trouver de nouveaux membres d'une famille de protéines dont on connaitrait pas la structure. La mise en place d'une série de descripteurs pour filtrer et ne garder que les meilleures séquences théoriques permet désormais de sélectionner de nouvelles protéines mutantes candidates pour la validation expérimentale. Complétée par des simulations de dynamique moléculaire, ce protocole permettra à notre équipe de tester d'autres protéines mutantes dans l'avenir. Ils pourront ainsi modifier des paramètres lors de la génération par CPD et s'appuyer sur des résultats expérimentaux pour les ajuster. Pour mieux aborder le design de protéine, il conviendrait d'incorporer des modifications importantes comme la flexibilité dans le squelette peptidique. Néanmoins, nous avons montré qu'avec un modèle assez

simple, nous avons obtenu des résultats très encourageants dans le domaine ambitieux du design de protéines.

L'ensemble des travaux de notre équipe, intégrant entre autres la paramétrisation du modèle de génération, l'utilisation des simulations de dynamique moléculaire, sera décrit dans un troisième article en cours de soumission en 2013 (Computational Protein Design : the Proteus Software and Selected Applications T. Simonson,T. Gaillard, D. Mignon, M. Schmidt am Busch, A. Lopes, N. Amara, S. Polydorides, A. Sedano, K. Druart, G. Archontis).

Protocoles, méthodes et annexes

Annexe A

Protocoles et méthodes expérimentaux

A.1 Techniques générales de biologie moléculaire, réactifs et tampons de purification

A.1.1 Souches de bactéries

Les différentes souches de bactéries ont été transformées par électroporation effectuée dans un appareil Gene Pulser (Biorad, Hercules, Californie) selon les recommandations du constructeur. Les bactéries électrocompétentes ont été préparées selon le protocole de Dower (1988). Les génotype des différentes souches utilisées dans cette étude sont répertoriées ci-dessous :

A.1.1.1 Préparation des cellules chimiocompétentes d'E. coli

Les bactéries provenant d'un aliquot de 45μL congelées à -80 $^\circ$C sont striées sur une boîte LB supplémenté en ampicilline et incubées à 37 $^\circ$C pendant une nuit. Une pré-culture de nuit est réalisée en ensemençant une colonie isolée de la boîte dans 3mM de milieu LB + antibiotique à 37 $^\circ$C. On ensemence 50mL de milieu LB + antibiotique par 1mL de pré-culture (au 1/50ième). La culture est incubée à 37 $^\circ$C sous agitation vive (190rpm) pendant 4h jusqu'à une DO_{650} comprise entre 0,4 et 0,6. Les étapes suivantes sont effectuées stérilement : les bactéries sont centrifugées à 3500xg pendant 15min puis

lavées avec 4mL de CaCl2 100mM, de nouveau centrifugées à 3500xg 15min à 4 ° C, puis repris dans 2mL de CaCl2 100mM (stérile) + 10 % glycérol à 4 ° C (1mL de CaCl2 100mM + 1mL de CaCl2 100mM + 20 % glycérol). Les cellules sont aliquotées par 45μL et congelées à -80 ° C.

A.1.2 Plasmides utilisés

A.1.3 Milieux et conditions de cultures

Les compositions des différents milieux de culture utilisés sont fournis ci-après pour 1L, les différent composants sont commandés chez Difco : Milieux liquides :

LB : 10g de tryptone, 5g d'extrait de levure, 5g de chlorure de sodium (NaCl)

2xTY : 16g tryptone, 10g d'extrait de levure, 5g de NaCl

Milieux solides :

LBAgar : Milieu LB + 0,2 % w/v d'agar

HTOP : 10g tryptone, 8g NaCl, 8g agar

A.1.4 Plasmides, préparations plasmidiques et caractérisation

Les minipréparations d'ADN plasmidiques ont été obtenues selon le protocole décrit par Birnboim et Doly (1979) ou par l'utilisation d'un protocole automatisé opéré par le robot Genesis (Tecan) en utilisant la méthode Nucleobond Spin (Macherey Nagel). Les systèmes QIAfiler (QIAGEN) ou Nucleobond (Macherey Nagel) ont permis d'obtenir des préparations d'ADN plasmidiques à grande échelle. Pour extraire l'ADN plasmidique nous avons utilisé le Kit QIAPrep Sip Miniprep (Qiagen) à partir de 2mL culture bactérienne en phase exponentielle de croissance. Pour les préparations d'ADN en grande quantité, nous avons utilisé le Kit QiaPrep sip Midiprep (Qiagen) à partir de culture de 100ml.

La caractérisation par électrophorèse d'ADN ont été réalisées par migration de gels à 1 % p/v d'agarose LE (Roche, Manheim, Allemagne) en présence de SYBR Green (Invi-

trogen). Les migrations sont réalisées dans un tampon TBE Tris 50mM, Borate 60mM, EDTA 1mM, pH 8,3. Le tampon de charge 6X (Bleu de Bromophénol 0.25 %, xylène cyanol FF 0,125 %, glycérol 30 %) est ajouté à l'ADN. Le marqueur de taille utilisé est le 1kb plus DNA Ladder (Invitrogen). L'ADN est révélé sous UV à 254nm. Sur un appareil UV2101PC (Shimadzu).

Les concentrations des préparations d'ADN plasmidique sont évaluées par l'acquisition de spectres d'absorption dans la gamme UV-visible sur un appareil UC2101PC (Shimadzu) ou ND-1000p (NanoDrop) en considérant qu'une unité d'absorbance à 260nm correspondant à une valeur approximative de 50μg/mL d'ADN double brin.

A.1.5 Clonage

Les manipulations des séquences d'ADN plasmidique ont été effectuées à l'aide du logiciel ApE (A plasmid Editor - Wayne Davis). Les oligonucléotides utilisés ont été commandés chez MWG Biotech AG (Ebersberg, Allemagne) et les désoxynucléotides proviennent d'Amerscham (Uppsala, Suède). Les enzymes de restrictions ont été commandées chez Roche (Manheim, Allemagne) ou New England Biolabs (Beverly, Massachussetts).

A.1.5.1 Amplification par PCR

L'amplification des fragments d'ADN par PCR ont été réalisés dans un volume réactionnel de 25μL contenant 5 ou 10 ng de matrice plasmidique ou 10 ng de matrice génomique, 50 pmoles de chacune des amorces nucléotidiques et un concentration finale de 200μM en chacun des désoxyribonucléotides (dATP, dCTP, dGTP, dTTP) ainsi que 0,4 unités de la polymérase Cloned Pfu (Stratagen) dans le tampon fourni avec la Taq DNA Polymerase. L'utilisation de la polymérase de Stratagene permet une relecture du fragment amplifié et donc une meilleure fidélité de cette biosynthèse. Les réactions de PCR sont réalisées en 30 cycles précédés d'une phase préalable de dénaturation à 95 $^\circ$C pendant 2 minutes. Les cycles comportent alors 15 secondes de dénaturation à 95 $^\circ$C, 30

secondes d'hybridation des oligonucléotides sur la matrice d'ADN (température en fonction des oligonucléotides) et une minute d'élongation à 72 ° C. Le programme s'achève par une phase d'élongation de 5 minutes à 72 ° C. La quantité de fragment synthétisé est alors contrôlée par dépôt de 5μL du mélange réactionnel sur gel d'agarose. Le mélange réactionnel est alors injecté sur un tamis moléculaire adapté à la purification de l'ADN (TSK Gel DNA, TOSOHAAS, Japon) (Schmitter et al., 1986). Les fractions de 320 μL correspondant aux fragments d'ADN sont concentrés par précipitation en présence d'éthanol (EtOH) à 66 %, d'acétate de sodium (NaAc) pH 4,6 à 0,3M et 10μg de Dextran DT40 par fraction. Ils sont ensuite hydrolysés par une quantité adaptée des enzymes de restriction avant d'être précipités par un mélange EtOH/NaAc. Les vecteurs de destination sont de la même façon hydrolysés par les mêmes enzymes de restriction avant d'être précipité par le même mélange EtOH/NaAc pH 4,6 et dissous dans l'eau. Les quantités d'ADN sont ensuite évaluées par électrophorèse afin de réaliser la ligature en présence d'un excès molaire de 7 à 10 de fragment à insérer par rapport au vecteur.

A.1.5.2 Ligation

Les ligatures ont été réalisées avec l'ADN ligase du phage T4 (GIBCO BRL) ou l'ADN ligase FastLink (TEBU Bio, Epicentre Technologies, Madison, Wisconsin). Lorsque l'ADN ligase du phage T4 est utilisée, un mélange réactionnel de 20-25 μL contenant 10 ng de vecteur, la quantité appropriée de fragment digéré à insérer, 200μM final d'ATP, 4mM final de DTT, 2,5 unité de ligase en présence de 50mM Tris-HCl pH 8,0, 10mM $MgCl_2$. Le mélange est alors laissé 15 minutes sur paillasse puis la ligase est inactivée en chauffant le mélange réactionnel pendant 10min à 70 ° C. Les réactions de ligatures sont alors hydrolysées par une enzyme de restriction dont un site est présent sur le vecteur de départ et pas sur le plasmide combiné attendu. Si nécessaire, une précipitation par un mélange EtOH/NaAc pH 4,6 est réalisée pour placer l'ADN dans un tampon compatible avec l'enzyme de restriction.

A.1.5.3 Transformation et ensemencement

1μL du mélange réactionnel est utilisé pour transformer les bactéries de la souche XL_1Blue. Suivant l'électroporation, les bactéries sont mélangées à 3 mL de milieu HTOP préchauffé à 40 °C avant d'être étalées sur des boîtes de Pétri contenant un milieu LBA-gar complémenté par un antibiotique adapté à la résistance du plasmide. Après une nuit de culture à 37 °C, 8 colonies par clonages sont repiquées sur des boîtes LB-Agar complémentées en antibiotique et cette procédure est répétée le lendemain afin d'assurer la clonalité des colonies. Les minipréparations d'ADN plasmidiques sont ensuite réalisées. L'insertion est vérifiée par analyse de restriction. Deux clones identifiés comme positifs servent à ensemencer 100 mL de milieu LB complémenté en l'antibiotique adapté afin d'extraire l'ADN plasmidique en plus grande quantité. La séquence de l'insert est systématiquement déterminée avant que le plasmide ne serve à l'expression de la protéine codée par le fragment inséré.

A.1.6 Séquençage de l'ADN

Les séquences des différents plasmides ont été vérifiées en utilisant un séquenceur automatique Long ReadIR 4200 (Li-Cor, Lincoln, Nebraska). Les gels de migration ainsi que le tampon de migration ont été préparés en accord avec les recommandations du constructeur pour réaliser des gels de 41 cm. Les fragments d'ADN obtenus par la méthode de terminaison de chaîne décrite par Sanger (1982) sont détectés par excitations de groupements fluorophores dans l'infrarouge (IRDye 700 et IRDye 800), greffés à l'extrémité 5' d'amorces nucléotidiques de 21 bases commandées chez MWG Biotech AG (Ebersberg, Allemagne). Les réactions de séquences par cyclage thermique ont été réalisées à l'aide de la méthode de séquençage d'Epicentre Technologies (TEBU Bio, Madison, Wisconsin) dans les appareils GeneAmp 2400 (Perkin Elmer, Wellesley, Maryland). Les séquences sont déterminées par le logiciel du séquenceur et les électrophorégrammes utilisés pour

lever les ambiguïtés. Les alignements de séquences d'ADN ont été réalisées à l'aide du programme DNAStrider (Marck, 1988).

A.1.7 Caractérisation des protéines par SDS-PAGE

La séparation des protéines en fonction de leurs tailles a été réalisée par électrophorèse en conditions dénaturantes (SDS-PAGE) à l'aide du système MiniProtean II de Biorad (Hercule, Californie). Les gels et tampon de migration ont été préparés selon Laemmli *et al.* (1970). Suivant la migration, les bandes de protéine sont révélées par coloration au bleu de Coomassie (Coomassie Brillant Blue de SIGMA) ou par coloration au nitrate d'argent (Silver Stain Plus de Biorad).

A.1.8 Production et purification à grande échelle des protéines

A.1.8.1 Principe de l'induction du gène d'intérêt

La surproduction du domaine src SH3 est réalisée dans la souche E. coli BL21(DE3). Cette souche dérivée de BL21 a intégré le phage λ DE3. Ce phage contient le gène codant l'ARN polymérase T7 en aval du promoteur lacUV5. Le gène src SH3 est cloné en aval d'un promoteur reconnu par l'ARN polymérase T7. La production est induite par ajout d'IPTG. L'IPTG (Isopropyl β-D-1-thiogalactopyranoside) est utilisé ici comme mimétique de l'allolactose, un métabolite du lactose qui se fixe sur le répresseur LacI levant ainsi la répression et déclenchant la transcription de l'opéron lac. L'ARN polymérase est alors produite et va permettre la production de la protéine d'intérêt présente sur le plasmide. Contrairement à l'allolactose, l'atome de soufre permet de créer une liaison chimique non hydrolysable par la cellule, ce qui évite la dégradation prématurée de l'inducteur.

A.1.8.2 Purification par chromatographie

Ces purifications ont été réalisées sur les systèmes FPLC Liquid Chromatography Controller LCC-500 (Amerscham Pharmacia Biotech, Upssala, Suède) ou bien par le système chromatographique de Biorad assisté par ordinateur BioLogic DuoFlow. Les colonnes de chromatographie Q-Hiload, S-Hiload, Mono Q, Mono S, Superdex 75, Superdex 200, Superdex 75HR, Superdex 200HR, proviennent du fabricant Amerscham Pharmacia Biotech. La résine IMAC TALON permettant l'accrochage spécifique des étiquettes histidine est commandée chez Clonetech. La protéase thrombine utilisée pour couper l'étiquette His-6 est commandée chez Sigma.

Les protéines purifiées ont été concentrées par centrifugation sur membrane en utilisant des unités de centrifugation Vivaspin (Vivascience AG, Hannovre, Allemagne) ou les unités AmiconUltra (Millipore). Les coefficients d'extinction molaires des protéines sont calculés à l'aide du logiciel Protparam (expasy). Les concentrations de protéine sont ainsi estimées par l'enregistrement de spectre d'absorption dans la gamme UV-visible sur un appareil UV2101PC ou ND-1000.

Le cassage des parois cellulaires est réalisé par sonication des bactéries resuspendues dans 50mL de tampon adapté (tampon de cassage). La composition des tampons utilisés au cours des différentes purifications sont indiquées ci-dessous :

Tampons de cassage :

– 10mM MOPS pH 6,7 ; 10mM β-mercapto-éthanol ; 0,1 mM EDTA ; 0,1mM PMSF

Tampons de séparation :

A 10mM MOPS pH 6,7 ; 10mM β-mercapto-éthanol ; 0,1 mM EDTA

B 10mM MOPS pH 6,7 ; 10mM β-mercapto-éthanol ; 0,1 mM EDTA ; 1M NaCl

A.1.8.3 Purification par interactions ioniques (chromatographie échangeuse d'ions)

La chromatographie à échange d'ions est un type de chromatographie en phase liquide permettant d'isoler une substance chargée électriquement d'un mélange de molécules chargées (liquide). Pour cela, on fait passer le mélange sur une phase stationnaire (solide) chargée déjà associée à des ions connus et on remplace ces ions par les ions/molécules chargées du mélange à séparer. Les protéines interfèrent avec un support (résine de silice ou autre, fonctionnalisé par des groupements ioniques positifs ou négatifs). Selon leur charge ionique (conditionné par le point isoélectrique (pI) et le pH du tampon), les protéines seront non fixées, retenues faiblement ou fortement, et donc séparées. Pour des purifications fines, ceci est réalisé dans des colonnes avec un débit de tampon qui peut varier, séparant ainsi les protéines selon un profil caractéristique.

A.1.9 Caractérisation de la séquence polypeptidique par spectrométrie de masse

Pour nous permettre de vérifier que les protéines purifiées et visibles sur gel polyacrylamide sont bien les domaines étudiés, nous avons eu recours à la spectrométrie de masse. Son principe réside dans la séparation en phase gazeuse de molécules chargées (ions) en fonction de leur rapport masse/charge (m/z). Nous digérons les échantillons par une enzyme la trypsine (clivage du côté C terminale au niveau des acides aminés basique Lysine et Arginine).

Les spectromètres de masse à haute résolution utilisés dans le laboratoire DCMR de l'Ecole Polytechnique :

- FT-ICR APEX III (Bruker) - Sources ESI, nano-ESI, MALDI
- Q-TOF Premier (Waters) - Sources ESI, nano-ESI

A.1.10 Étude des domaines SH3-1CKA sauvage et mutant Gen1-*old*

A.1.10.1 Clonage

Les séquences d'ADN codant pour les protéines SH3-1CKA sauvage et mutante Gen1-*old*, ont été commandées chez Epoch Biolabs, dont les codons y ont été optimisé. Le gènes d'intérêt a été cloné dans un plasmide pBluescript II SK(-) entre les sites de restriction NcoI et XhoI. Un Tag-6His a été rajouté en amont des gènes.

Puis le gène a été cloné dans pET15b lpa entre NcoI et XhoI. Au vue des échecs de surexpression, nous avons tenté ensuite de rallonger le Tag à 10 histidines par mutagénèse dirigée. La surexpression étant encore moins probante, nous avons donc décidé de supprimer le Tag et de cloner le gène dans pET3a entre les sites de restriction NdeI et BamHI.

A.1.10.2 Production et purification

Nous avons réussi à déterminer les conditions de surexpression pour la protéine sauvage SH3-1CKA et la protéine mutante Gen1-*old*. Voici le détail :

- transformation de 0,8μL maxi-préparation du plasmide SH3-1CKA dans la souche BL21(DE3)
- ensemencement de 5μL dans préculture de 5mL en milieu LB + 100mM ampicilline, à 37 $^\circ$C sur la nuit
- le matin, ensemencement à 0,05DO/mL dans 50mM de milieu LB+amp (préchauffé à 37 $^\circ$C)
- à 0,6DO/mL , ajout de 50mM IPTG
- 3h à 37degreeC

A.2 Étude structurale par Dichroïsme circulaire et calorimétrie

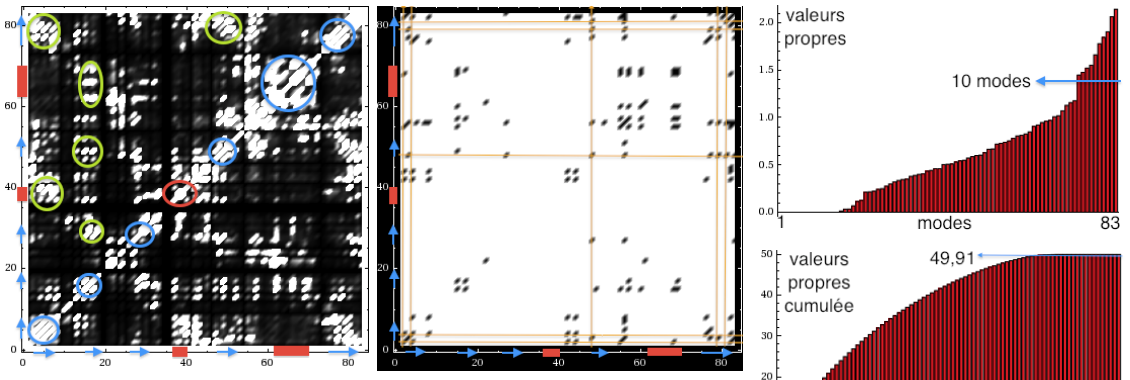
Les échantillons passés en dichroïsme circulaire étaient concentrés à 20μL dans un tampon NaH_2PO_4 à 20mM, pour un spectre CD entre 185nm et 260nm. Les échantillons pour la calorimétrie étaient concentrés à 0,5 mg/mL dans un tampon Hépès 10mM pH7,5 + EDTA 0,1mM.

A.3 Étude structurale par RMN

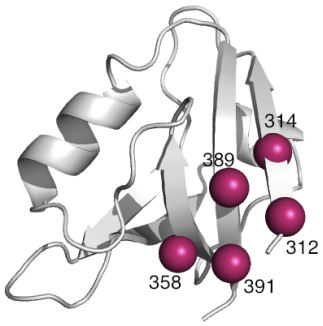
Nous avons réalisé nos cultures en milieu minimum M9 avec supplément de chlorure d'ammonium $^{15}NH_4Cl$ (Sambrook et al. 1989). Les échantillons étaient de 380μL avec 80mM NaCl, 20mM Hépès pH7,5 , 0,1mM EDTA, et 10% D_2O.

Annexe B

Annexes bioinformatiques

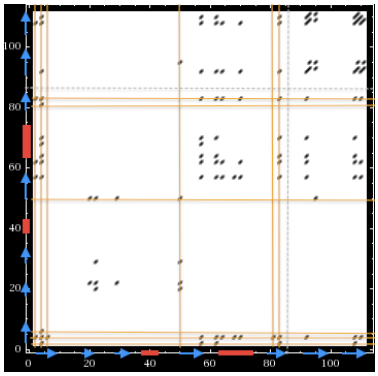


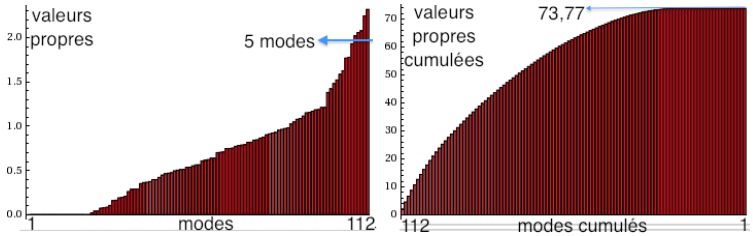
Trace matrice corrélation entière	49.9077
Trace modes 1 à 10 superposés	17.3846 (34.8%)
Trace modes 1 à 10 seuillés à 25%	15.335 (30.7%)



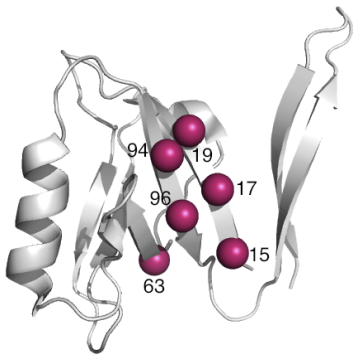
Nb séquences	Pourcentage	Instances motif
860	8,6%	DLDHD
836	8,4%	HLDHD
589	5,9%	DLLHW
408	4,1%	HLLHW
372	3,7%	WADHD
302	3,0%	HLDWD
265	2,7%	DLHHD
218	2,2%	WALHW
191	1,9%	HLHHD
181	1,8%	DLDWD
5778	57,7%	Reste
10000	100%	All séq.

Figure B.1 – Analyse des covariances pour la structure 1BE9 du domaine PDZ. (En haut à gauche) Matrice de covariance complète et matrice correspondant aux 10 premiers modes propres superposés avec un seuil de débruitage de 25%. (En haut à droite) Distribution des valeurs propres. (Au milieu à gauche) Traces (sommes des valeurs propres) des différentes matrices de covariance. (Au milieu à droite) Distribution des valeurs propres cumulées. (En bas à gauche) Structure 3D avec le motif : 312-314-358-389-391 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.



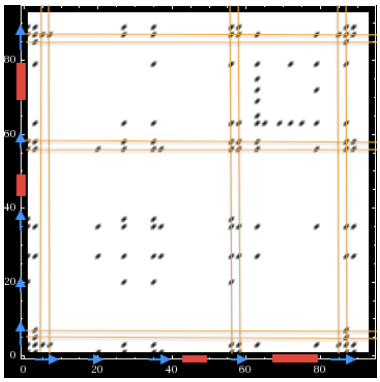


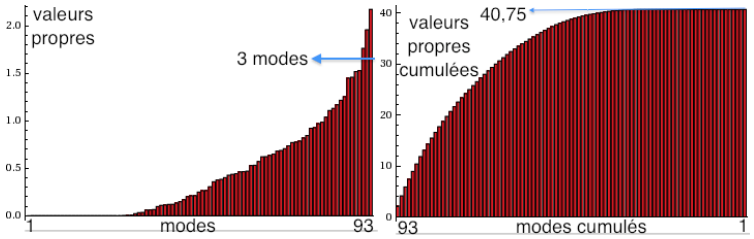
Trace matrice corrélation entière	73.7717
Trace modes 1 à 5 superposés	10.7178 (14.5%)
Trace modes 1 à 5 seuillés à 25%	7.85123 (10.6%)



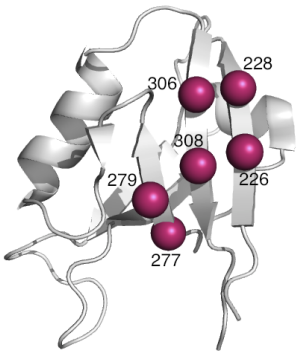
Nb séquences	Pourcentage	Instances motif
580	5,8%	DDLWHL
565	5,7%	DDLWHH
476	4,8%	DHDWHW
286	2,9%	HDLWHL
276	2,8%	DDDWHH
270	2,7%	WLLWHW
265	2,7%	WHDWHW
210	2,1%	WLDWHW
210	2,1%	WDLWHH
197	2,0%	DLLWHW
6665	66,7%	Reste
10000	100%	All séq.

Figure B.2 – Analyse des covariances pour la structure 1QAU du domaine PDZ. (En haut à gauche) Matrice de covariance correspondant aux 5 premiers modes propres superposés avec un seuil de débruitage de 25%. (En haut à droite) Distribution des valeurs propres, et des valeurs propres cumulées. Et traces des différentes matrices de covariance. (En bas à gauche) Structure 3D avec le motif : 15-17-19-63-94-96 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.



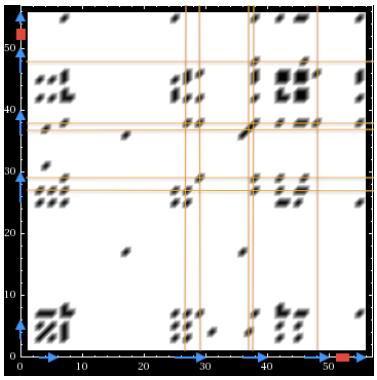


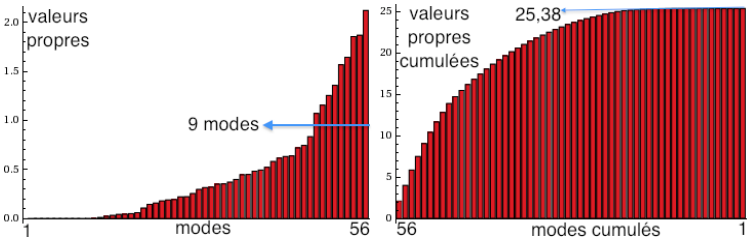
Trace matrice corrélation entière	40.7471
Trace modes 1 à 3 superposés	5.89154 (14.5%)
Trace modes 1 à 3 seuillés à 18%	4.65346 (11.4%)



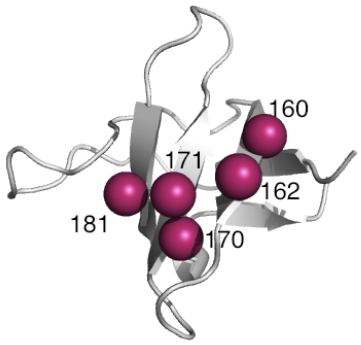
Nb séquences	Pourcentage	Instances motif
923	9.2%	DDDWHH
748	7.5%	DDWWHH
335	3.4%	DDDHHW
274	2.7%	DDDWHW
250	2.5%	DDDDHH
223	2.2%	DDDHHH
221	2.2%	DDWWHW
211	2.1%	DLWWWH
195	2.0%	DLDWWH
176	1.8%	HDDWWW
6444	64.4%	Reste
10000	100%	All séq.

Figure B.3 – Analyse des covariances pour la structure 2FE5 du domaine PDZ. (En haut à gauche) Matrice de covariance correspondant aux 3 premiers modes propres superposés avec un seuil de débruitage de 18%. (En haut à droite) Distribution des valeurs propres, et des valeurs propres cumulées. Et traces des différentes matrices de covariance. (En bas à gauche) Structure 3D avec le motif : 226-228-277-279-306-308 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.



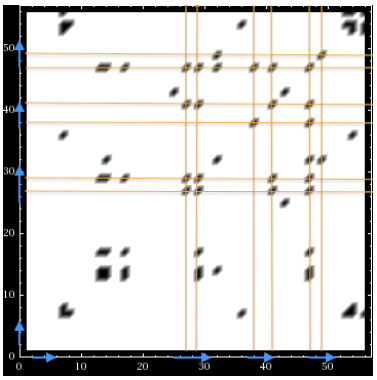


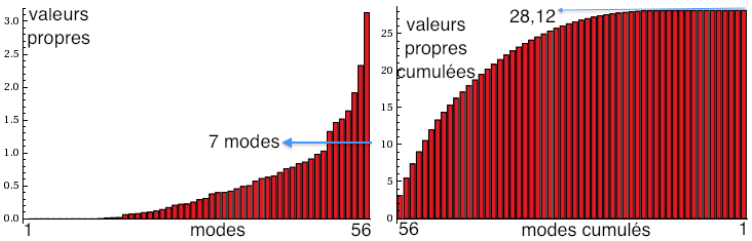
Trace matrice corrélation entière	25.3841
Trace modes 1 à 9 superposés	13.8847 (54.7%)
Trace modes 1 à 9 seuillés à 25%	12.1833 (48%)



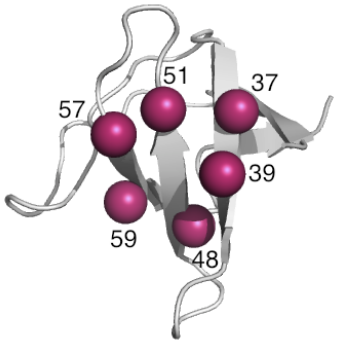
Nb séquences	Pourcentage	Instances motif
1185	11.9%	WLWWL
1161	11.6%	WLWLL
779	7.8%	WLWDH
494	4.9%	WLWDL
450	4.5%	HLWLL
396	4%	WLWHL
380	3.8%	DLWWL
321	3.2%	WHWDH
275	2.8%	HHWDH
238	2.4%	HLWWL
4321	43.1%	Reste
10000	100%	All séq.

Figure B.4 – Analyse des covariances pour la structure 1CKA du domaine SH3. (En haut à gauche) Matrice de covariance correspondant aux 9 premiers modes propres superposés avec un seuil de débruitage de 25%. (En haut à droite) Distribution des valeurs propres, et des valeurs propres cumulées. Et traces des différentes matrices de covariance. (En bas à gauche) Structure 3D avec le motif : 160-162-170-171-181 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.



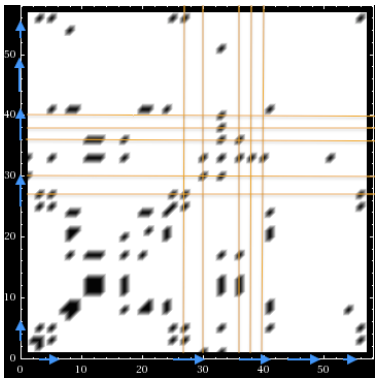


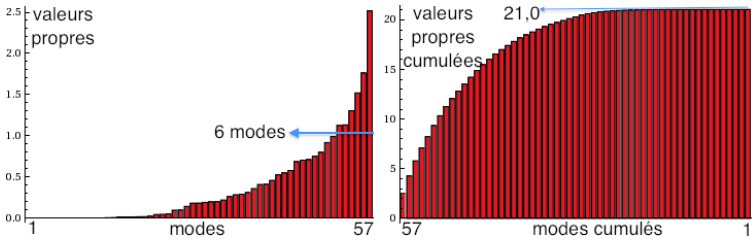
Trace matrice corrélation entière	28.12
Trace modes 1 à 7 superposés	13.32 (47.4%)
Trace modes 1 à 7 seuillés à 38%	10.04 (35.7%)



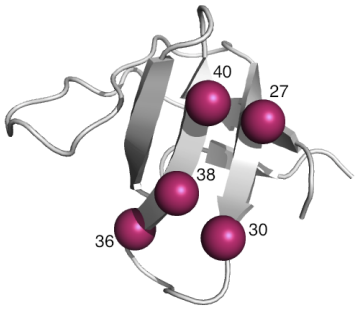
Nb séquences	Pourcentage	Instances motif
1100	11%	DWHHAD
1083	10.8%	DWHHAH
809	8.1%	DWHHDD
475	4.8%	DLHHWD
463	4.6%	DWHHDH
363	3.6%	WWLDDH
345	3.5%	WWHDDH
305	3.5%	DLHHWH
268	2.7%	WWHDDD
261	2.6%	WWLDDD
4528	45.3%	Reste
10000	100%	All séq.

Figure B.5 – Analyse des covariances pour la structure 1CSK du domaine SH3. (En haut à gauche) Matrice de covariance correspondant aux 7 premiers modes propres superposés avec un seuil de débruitage de 38%. (En haut à droite) Distribution des valeurs propres, et des valeurs propres cumulées. Et traces des différentes matrices de covariance. (En bas à gauche) Structure 3D avec le motif : 37-39-48-51-57-59 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.





Trace matrice corrélation entière	21.0439
Trace modes 1 à 6 superposés	9.35565 (44.5%)
Trace modes 1 à 6 seuillés à 10%	9.00356 (42.8%)



Nb séquences	Pourcentage	Instances motif
2351	23.5%	HLWWD
1691	16.9%	HHWWD
640	6.4%	HHDWD
602	6.0%	HHHWD
555	5.6%	DLWWD
512	5.51%	DHWWD
439	4.4%	HLWLL
412	4.1%	HLHWD
387	3.9%	HLDWD
177	1.8%	ALWWD
2234	22.3%	Reste
10000	100%	All séq.

Figure B.6 – Analyse des covariances pour la structure 1UTI du domaine SH3. (En haut à gauche) Matrice de covariance correspondant aux 6 premiers modes propres superposés avec un seuil de débruitage de 10%. (En haut à droite) Distribution des valeurs propres, et des valeurs propres cumulées. Et traces des différentes matrices de covariance. (En bas à gauche) Structure 3D avec le motif : 27-30-36-38-40 représenté par des boules roses. (En bas à droite) Nombre de séquences et pourcentage par rapport au nombre total de séquences considérées correspondant au motif retenu.

Protéines homologues sorties sans groupe (redondants) :					
A8E0R9|GRIP2_XENLA		Q5RBI7|PDZD7_PONAB		Q92796|DLG3_HUMAN	
O14745|NHERF_HUMAN		Q5RCF7|PDZD1_PONAB		Q93646|SNT1_CAEEL	
O14907|TX1B3_HUMAN		Q5T2W1|PDZD1_HUMAN		Q96JH8|K1849_HUMAN	
O14910|LIN7A_HUMAN		Q5TCQ9|MAGI3_HUMAN		Q96NW7|LRRC7_HUMAN	
O15018|PDZD2_HUMAN		Q61235|SNTB2_MOUSE		Q96QZ7|MAGI1_HUMAN	
O35867|NEB1_RAT		Q62108|DLG4_MOUSE		Q96RT1|LAP2_HUMAN	
O35889|AFAD_RAT		Q62696|DLG1_RAT		Q99L88|SNTB1_MOUSE	
O55164|MPDZ_RAT		Q62936|DLG3_RAT		Q99NH2|PARD3_MOUSE	
O62683|ZO3_CANFA		Q63622|DLG2_RAT		Q9DBG9|TX1B3_MOUSE	
O75970|MPDZ_HUMAN		Q63ZW7|INADL_MOUSE		Q9EQJ9|MAGI3_MOUSE	
O88382|MAGI2_RAT		Q64512|PTN13_MOUSE		Q9ES64|USH1C_MOUSE	
O88951|LIN7B_MOUSE		Q68DX3|FRPD2_HUMAN		Q9H5P4|PDZD7_HUMAN	
P31007|DLG1_DROME		Q69Z89|K1849_MOUSE		Q9HAP6|LIN7B_HUMAN	
P31016|DLG4_RAT		Q69ZH9|RHG23_MOUSE		Q9JHL1|NHRF2_MOUSE	
P55196|AFAD_HUMAN		Q69ZS0|PZRN3_MOUSE		Q9JIL4|PDZD1_MOUSE	
P57105|SYJ2B_HUMAN		Q6R005|DLG4_DANRE		Q9JJ40|PDZD1_RAT	
P68907|PZRN3_RAT		Q6RHR9|MAGI1_MOUSE		Q9JK71|MAGI3_RAT	
P70175|DLG3_MOUSE		Q6ZWJ1|STXB4_HUMAN		Q9P202|WHRN_HUMAN	
P70587|LRRC7_RAT		Q7KRY7|LAP4_DROME		Q9QZR8|PDZD2_RAT	
P78352|DLG4_HUMAN		Q80TE7|LRRC7_MOUSE		Q9ULJ8|NEB1_HUMAN	
Q12923|PTN13_HUMAN		Q80TH2|LAP2_MOUSE		Q9UPQ7|PZRN3_HUMAN	
Q13425|SNTB2_HUMAN		Q80U72|LAP4_MOUSE		Q9WV89|STXB4_MOUSE	
Q13884|SNTB1_HUMAN		Q80VW5|WHRN_MOUSE		Q9WVJ4|SYJ2B_RAT	
Q14160|LAP4_HUMAN		Q810W9|WHRN_RAT		Q9WVQ1|MAGI2_MOUSE	
Q15599|NHRF2_HUMAN		Q811D0|DLG1_MOUSE		Q9Y4G8|RPGF2_HUMAN	
Q28619|NHERF_RABIT		Q865P3|PDZD1_RABIT		Q9Z250|LIN7A_RAT	
Q2KIB6|LIN7B_BOVIN		Q86UL8|MAGI2_HUMAN		Q9Z252|LIN7B_RAT	
Q3SZK8|NHERF_BOVIN		Q8JZS0|LIN7A_MOUSE		Q9Z340|PARD3_RAT	
Q3T0X8|PDZD1_BOVIN		Q8NI35|INADL_HUMAN		Q28C55|DLG1_XENTR	
Q4L1J4|MAGI1_RAT		Q8SQG9|NHRF2_RABIT		Q3T0C9|SYJ2B_BOVIN	
Q4R6G4|NHERF_MACFA		Q8TEU7|RPGF6_HUMAN		Q4H4B6|SCRIB_DANRE	
Q5PYH5|DLG1L_BRARE		Q8VBX6|MPDZ_MOUSE		Q792I0|LIN7C_RAT	
Q5PYH6|DLG1_BRARE		Q91XM9|DLG2_MOUSE		Q9D6K5|SYJ2B_MOUSE	
Q5PYH7|DLG2_DANRE		Q920G2|NHRF2_RAT			
Protéines homologues sorties qu'avec les groupes (redondants) :					
A1A5G4|MAGI3_XENTR		Q5RAA5|LIN7C_PONAB		Q8TEW0|PARD3_HUMAN	
O35274|NEB2_RAT		Q5RD32|GOPC_PONAB		Q8TEW8|PAR3L_HUMAN	
O54824|IL16_MOUSE		Q5ZM14|NHERF_CHICK		Q96SB3|NEB2_HUMAN	
Q12959|DLG1_HUMAN		Q6R891|NEB2_MOUSE		Q9QXY1|ZO3_MOUSE	
Q15700|DLG2_HUMAN		Q6ZMN7|PZRN4_HUMAN		Q9QZQ1|AFAD_MOUSE	
Q5F425|LIN7C_CHICK		Q8BH60|GOPC_MOUSE			
Q5F488|MAGI3_CHICK		Q8TDM6|DLG5_HUMAN			
Liste tous les protéines homologues non redondantes :					
A8E0R9	G45	Q68DX3	Q93646		
G10	G47	Q69Z89	Q96JH8		
G11	G51	Q69ZH9	Q96NW7		
G12	G55	Q6ZMN7	Q96RT1		
G15	G56	Q7KRY7	Q9ES64		
G20	G57	Q80TE7	Q9JHL1		
G21	G58	Q80TH2	Q9QXY1		
G3	G6	Q865P3	Q9Y4G8		
G35	G62	Q8SQG9			
G36	G64	Q8TDM6			
G42	G66	Q8TEU7			
G43	G67	Q920G2			

Figure B.7 – Liste des protéines homologues trouvées pour la protéine PDZ-1BE9.

Protéines homologues sorties sans groupe (redondants) :		
A1A5G4|MAGI3_XENTR	Q2KIB6|LIN7B_BOVIN	Q8TEW8|PAR3L_HUMAN
A8E0R9|GRIP2_XENLA	Q4L1J4|MAGI1_RAT	Q8VBX6|MPDZ_MOUSE
O14910|LIN7A_HUMAN	Q5F425|LIN7C_CHICK	Q91XM9|DLG2_MOUSE
O15018|PDZD2_HUMAN	Q5F488|MAGI3_CHICK	Q925T6|GRIP1_MOUSE
O35274|NEB2_RAT	Q5PYH5|DLG1L_BRARE	Q92796|DLG3_HUMAN
O35867|NEB1_RAT	Q5PYH6|DLG1_BRARE	Q96JH8|K1849_HUMAN
O55164|MPDZ_RAT	Q5PYH7|DLG2_DANRE	Q96NW7|LRRC7_HUMAN
O62666|IL16_PANTR	Q5RAA5|LIN7C_PONAB	Q96QZ7|MAGI1_HUMAN
O62674|IL16_CERAE	Q5TCQ9|MAGI3_HUMAN	Q96RT1|LAP2_HUMAN
O62675|IL16_MACMU	Q5ZM14|NHERF_CHICK	Q96SB3|NEB2_HUMAN
O62676|IL16_MACFA	Q62108|DLG4_MOUSE	Q99NH2|PARD3_MOUSE
O62677|IL16_SAISC	Q62696|DLG1_RAT	Q9C0E4|GRIP2_HUMAN
O62678|IL16_AOTTR	Q62936|DLG3_RAT	Q9CSB4|PAR3L_MOUSE
O75970|MPDZ_HUMAN	Q63622|DLG2_RAT	Q9EQJ9|MAGI3_MOUSE
O88382|MAGI2_RAT	Q63ZW7|INADL_MOUSE	Q9HAP6|LIN7B_HUMAN
O88951|LIN7B_MOUSE	Q69Z89|K1849_MOUSE	Q9JK71|MAGI3_RAT
P31007|DLG1_DROME	Q6R005|DLG4_DANRE	Q9NB04|PATJ_DROME
P31016|DLG4_RAT	Q6R891|NEB2_MOUSE	Q9PU36|PCLO_CHICK
P57105|SYJ2B_HUMAN	Q6RHR9|MAGI1_MOUSE	Q9QYX7|PCLO_MOUSE
P70175|DLG3_MOUSE	Q7KRY7|LAP4_DROME	Q9QZR8|PDZD2_RAT
P70587|LRRC7_RAT	Q80TE7|LRRC7_MOUSE	Q9ULJ8|NEB1_HUMAN
P78352|DLG4_HUMAN	Q80TH2|LAP2_MOUSE	Q9WVJ4|SYJ2B_RAT
P97879|GRIP1_RAT	Q80U72|LAP4_MOUSE	Q9WVQ1|MAGI2_MOUSE
Q12923|PTN13_HUMAN	Q811D0|DLG1_MOUSE	Q9Y3R0|GRIP1_HUMAN
Q12959|DLG1_HUMAN	Q86UL8|MAGI2_HUMAN	Q9Z250|LIN7A_RAT
Q14005|IL16_HUMAN	Q8JZS0|LIN7A_MOUSE	Q9Z252|LIN7B_RAT
Q14160|LAP4_HUMAN	Q8NI35|INADL_HUMAN	Q9Z340|PARD3_RAT
Q15700|DLG2_HUMAN	Q8TEW0|PARD3_HUMAN	

Protéines homologues sorties qu'avec les groupes (redondants) :
P51140|DSH_DROME
Q0V8R5|IL16_BOVIN
Q4KL35|MAGIX_MOUSE
Q9JKS6|PCLO_RAT
Q9WTW1|GRIP2_RAT

Liste des protéines homologues non redondantes			
A8E0R9	G41	G74	Q80TE7
G10	G42	G8	Q80TH2
G11	G47	O62677	Q96JH8
G15	G51	P31007	Q96NW7
g20	G55	P51140	Q96RT1
G20	G62	P70587	Q9NB04
G21	G66	Q4KL35	Q9PU36
G3	G70	Q5ZM14	Q9QYX7
G35	G71	Q69Z89	
G39	G72	Q7KRY7	

Figure B.8 – Liste des protéines homologues trouvées pour la protéine PDZ-2FE5.

Protéines homologues sorties sans groupe (redondants) :			
O19132|NOS1_RABIT	Q29498|NOS1_SHEEP		
P29475|NOS1_HUMAN	Q91XM9|DLG2_MOUSE		
P29476|NOS1_RAT	Q9Z0J4|NOS1_MOUSE		
Protéines homologues sorties qu'avec les groupes (redondants) :			
Q15700|DLG2_HUMAN	Q64512|PTN13_MOUSE		
Q63622|DLG2_RAT	Q86UT5|PDZD3_HUMAN		
Liste tous les protéines homologues non redondantes :			
G4	G49		
G42	Q86UT5		
G47			

Figure B.9 – Liste des protéines homologues trouvées pour la protéine PDZ-1QAU.

Protéines homologues sorties sans groupe (redondants) :		
P05433|GAGC_AVISC	P47941|CRKL_MOUSE	Q64010|CRK_MOUSE
P32577|CSK_RAT	P87378|CRK_XENLA	Q6DCZ7|FBP1L_XENLA
P41239|CSK_CHICK	Q04929|CRK_CHICK	Q6GUF4|FBP1L_XENTR
P41240|CSK_HUMAN	Q0VBZ0|CSK_BOVIN	Q8K012|FBP1L_MOUSE
P41241|CSK_MOUSE	Q2HWF0|FBP1L_RAT	Q9XYM0|CRK_DROME
P46108|CRK_HUMAN	Q5T0N5|FBP1L_HUMAN	
P46109|CRKL_HUMAN	Q63768|CRK_RAT	
Protéines homologues sorties qu'avec groupes (redondantes) :		
O15034|RIMB2_HUMAN	Q8QFX1|RIMB2_CHICK	
P29355|SEM5_CAEEL	Q9JIR1|RIMB2_RAT	
Q80U40|RIMB2_MOUSE		
Liste de toutes les protéines homologues non redondantes :		
G29	G9	
G31	P29355	
G32	P87378	
G50	Q9XYM0	

Figure B.10 – Liste des protéines homologues trouvées pour la protéine SH3-1CSK.

Protéines homologues sorties sans groupe (redondants) :		
A0JNJ1|STAC_BOVIN	P87379|GRB2A_XENLA	Q6CVA8|HSE1_KLULA
A4RF61|HSE1_MAGGR	Q06449|PIN3_YEAST	Q6WKZ7|NOSTN_MOUSE
O42287|ITSN1_XENLA	Q13588|GRAP_HUMAN	Q7S6J4|HSE1_NEUCR
O43125|CSH3_SCHPO	Q13882|PTK6_HUMAN	Q7TNR9|ARHG4_MOUSE
O43307|ARHG9_HUMAN	Q14155|ARHG7_HUMAN	Q8IVI9|NOSTN_HUMAN
O55043|ARHG7_RAT	Q15052|ARHG6_HUMAN	Q8K4I3|ARHG6_MOUSE
O75791|GRAP2_HUMAN	Q15811|ITSN1_HUMAN	Q8WUF5|ASPP_HUMAN
O89100|GRAP2_MOUSE	Q2KJB5|NOSTN_BOVIN	Q92783|STAM1_HUMAN
O93436|STAM2_CHICK	Q3UTH8|ARHG9_MOUSE	Q96KQ4|ASPP1_HUMAN
P10569|MYSC_ACACA	Q45FX5|VAV_CAEEL	Q96N96|SPT13_HUMAN
P15498|VAV_HUMAN	Q58DL7|ARHG9_BOVIN	Q99469|STAC_HUMAN
P27870|VAV_MOUSE	Q5DU57|SPT13_MOUSE	Q9CX99|GRAP_MOUSE
P29355|SEM5_CAEEL	Q5I0D6|NOSTN_RAT	Q9ES28|ARHG7_MOUSE
P43603|LSB3_YEAST	Q5I1X5|ASPP_MOUSE	Q9NR80|ARHG4_HUMAN
P52735|VAV2_HUMAN	Q5RDK0|ARHG9_PONAB	Q9QX73|ARHG9_RAT
P53281|LSB1_YEAST	Q5XXR3|ARHG6_RAT	Q9R144|ANM2_MOUSE
P54100|VAV_RAT	Q5ZLR6|ARHG6_CHICK	Q9WVE9|ITSN1_RAT
P55345|ANM2_HUMAN	Q60992|VAV2_MOUSE	Q9Z0R4|ITSN1_MOUSE
P70297|STAM1_MOUSE	Q62415|ASPP1_MOUSE	

Protéines homologues sorties qu'avec les groupes (redondants) :		
O74749|HSE1_SCHPO	Q2GT05|HSE1_CHAGB	Q9NZM3|ITSN2_HUMAN
O88811|STAM2_MOUSE	Q3SZ01|SH3Y1_BOVIN	Q9UKW4|VAV3_HUMAN
P05433|GAGC_AVISC	Q5BBL4|HSE1_EMENI	Q9URW6|YIE2_SCHPO
P12931|SRC_HUMAN	Q5R4J7|GRB2_PONAB	Q9Z0R6|ITSN2_MOUSE
P13115|SRC1_XENLA	Q60631|GRB2_MOUSE	
Q07883|GRB2_CHICK	Q66II3|GRB2_XENTR	
Q0U6X7|HSE1_PHANO	Q68F99|STAC3_XENTR	
Q1LYG0|SPD2A_DANRE	Q6GPJ9|GRB2B_XENLA	

Liste des protéines homologues non redondantes :			
G13	G38	G76	Q45FX5
G14	G40	O43125	Q68F99
G17	G44	O74749	Q6CVA8
G18	G46	P10569	Q9UKW4
G2	G5	P29355	Q9URW6
G23	G52	P43603	
G27	G54	P53281	
G31	G60	Q06449	
G33	G65	Q13882	
G34	G69	Q1LYG0	
G37	G75	Q3SZ01	

Figure B.11 – Liste des protéines homologues trouvées pour la protéine SH3-1UTI.

```
G1:A0JNB0 A1Y2K1 P06241 P39688 Q05876 Q62844 P13406
G2:A0JNJ1 Q99469
G3:A1A5G4 Q5TCQ9 Q9EQJ9 Q9JK71 Q5F488
G4:A3D7K1 A6WRG3
G5:A4RF61 Q2GT05 Q7S6J4
G6:O14745 Q4R6G4 Q28619 Q3SZK8
G7:O14907 Q9DBG9
G8:O15018 Q9QZR8
G9:O15034 Q80U40 Q8QFX1 Q9JIR1
G10:O35274 Q6R891 Q96SB3
G11:O35867 Q9ULJ8
G12:O35889 Q9QZQ1 P55196
G13:O42287 Q9Z0R4 Q15811 Q9WVE9
G14:O55043 Q9ES28 Q14155
G15:O55164 O75970 Q8VBX6
G16:O75044 Q91Z67 O43295 Q812A2 Q7Z6B7 Q91Z69
G17:O75791 O89100
G18:O75886 O88811 Q5XHY7 O93436
G19:O75962 Q0KL02
G20:O88382 Q86UL8 Q9WVQ1
G21:O88951 Q9Z252 O14910 Q8JZS0 Q9Z250 Q5F425 Q5RAA5 Q792I0 Q2KIB6 Q9HAP6
G22:O89032 Q5TCZ1
G23:P00523 P00525 P14084 P14085 P15054 P12931 Q9WUD9 P13115 P13116 P00526
P25020 P63185 P31693
G24:P00527 P09324 Q28923 Q04736 P10936 P07947
G25:P07948 P25911 Q07014
G26:P09769 Q02977
G27:P15498 P27870 P54100
G28:P24604 P42680
G29:P32577 P41240 P41241 Q0VBZ0 P41239
G30:P42686 P42690
G31:P46108 Q63768 Q64010 P05433 Q04929
G32:P46109 P47941 Q5U2U2
G33:P52735 Q60992
G34:P55345 Q9R144
G35:P57105 Q3T0C9 Q9WVJ4 Q9D6K5
G36:P68907 Q69ZS0 Q9UPQ7
G37:P70297 Q92783
G38:P87379 Q6GPJ9 Q5R4J7 Q60631 Q66II3 Q07883 P62994 Q9NYB9 P62484
G39:P97879 Q925T6 Q9Y3R0
G40:Q0U6X7 Q4WHP5 Q5BBL4
G41:Q0V8R5 O62674 O62675 O62676 O62666 Q14005 O62678
G42:Q12923 Q64512
G43:Q13425 Q61235
G44:Q13588 Q9CX99 P62484
G45:Q13884 Q99L88
G46:Q15052 Q5XXR3 Q8K4I3 Q5ZLR6
G47:Q15700 P70175 Q62936 Q92796 Q5PYH7 Q63622 Q91XM9
G49:Q29498 O19132 P29475 P29476 Q9Z0J4
G50:Q2HWF0 Q8K012 Q5T0N5 Q6GUF4 Q6DCZ7
```

Figure B.12 – Liste des identifiants des protéines homologues entre elles pour les domaines SH3 et PDZ.

```
G51:Q4L1J4 Q96QZ7 Q6RHR9
G52:Q5DU57 Q96N96
G53:Q5FVW6 Q5U597
G54:Q5I1X5 Q8WUF5
G55:Q5PYH6 Q12959 Q811D0 Q62696 P31016 P78352 Q62108 Q5PYH5 Q6R005 Q28C55
G56:Q5RBI7 Q9H5P4
G57:Q5RCF7 Q5T2W1
G58:Q5RD32 Q8BH60
G59:Q5U228 Q1LVQ2 Q921I6 Q9JJS5 Q9P0V3
G60:Q62415 Q96KQ4
G61:Q62422 Q8MJ50 Q8MJ49
G62:Q63ZW7 Q8NI35
G63:Q6TGW5 Q6XJU9
G64:Q6ZWJ1 Q9WV89
G65:Q7TNR9 Q9NR80
G66:Q80U72 Q4H4B6 Q14160
G67:Q80VW5 Q810W9 Q9P202
G68:Q8CBW3 Q9QZM5
G69:Q8IVI9 Q6WKZ7 Q2KJB5 Q5I0D6
G70:Q8TEW0 Q99NH2 Q9Z340
G71:Q8TEW8 Q9CSB4
G72:Q9C0E4 Q9WTW1
G73:Q9JIL4 Q9JJ40
G74:Q9JKS6 Q9QYX7
G75:Q9NZM3 Q9Z0R6
G76:Q9QX73 O43307 Q3UTH8 Q58DL7 Q5RDK0
```

Figure B.13 – Liste des identifiants des protéines homologues entre elles pour les domaines SH3 et PDZ.

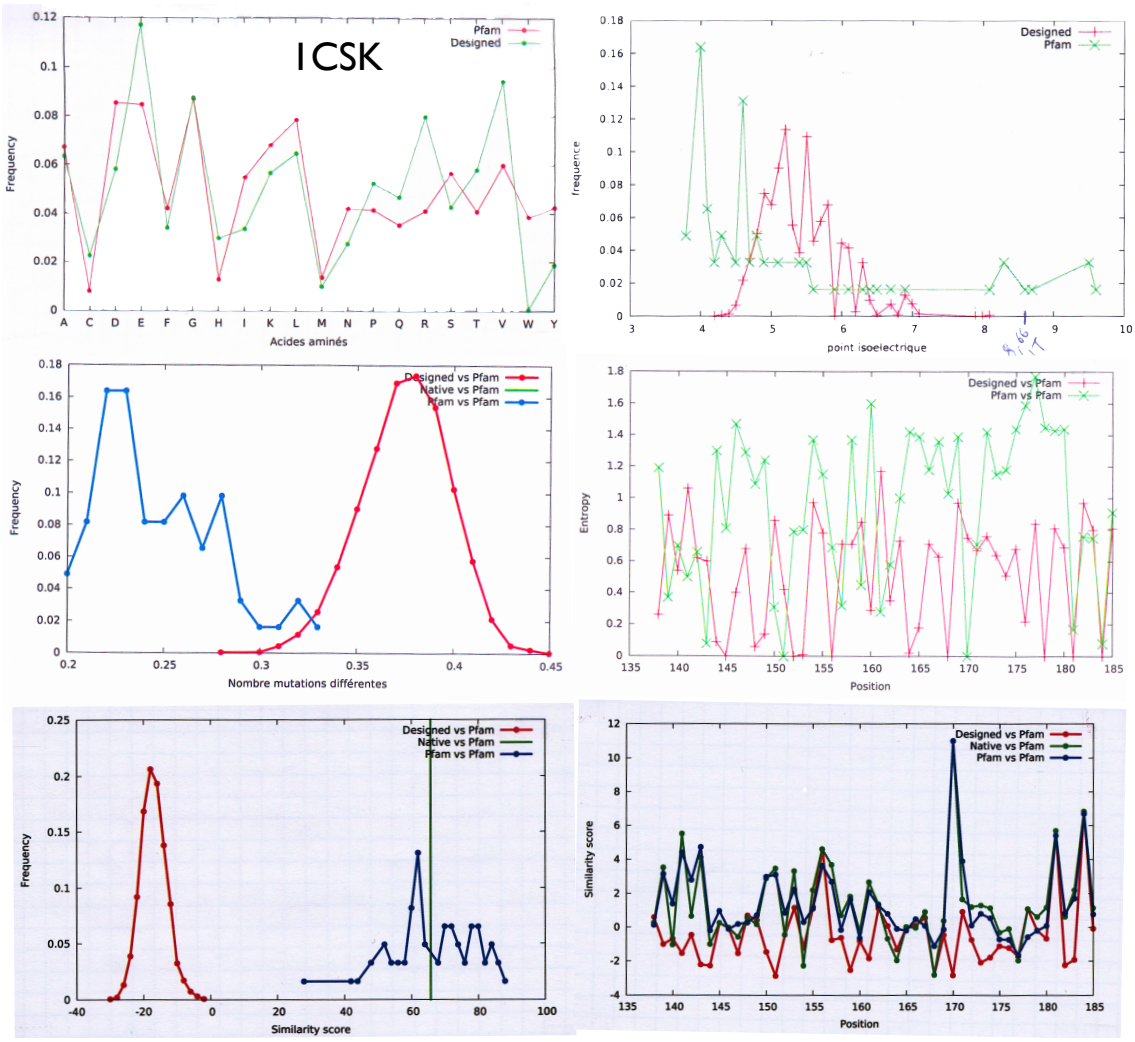


Figure B.14 – Descripteurs appliqués sur les séquences théoriques générées sur la structure 1CSK.

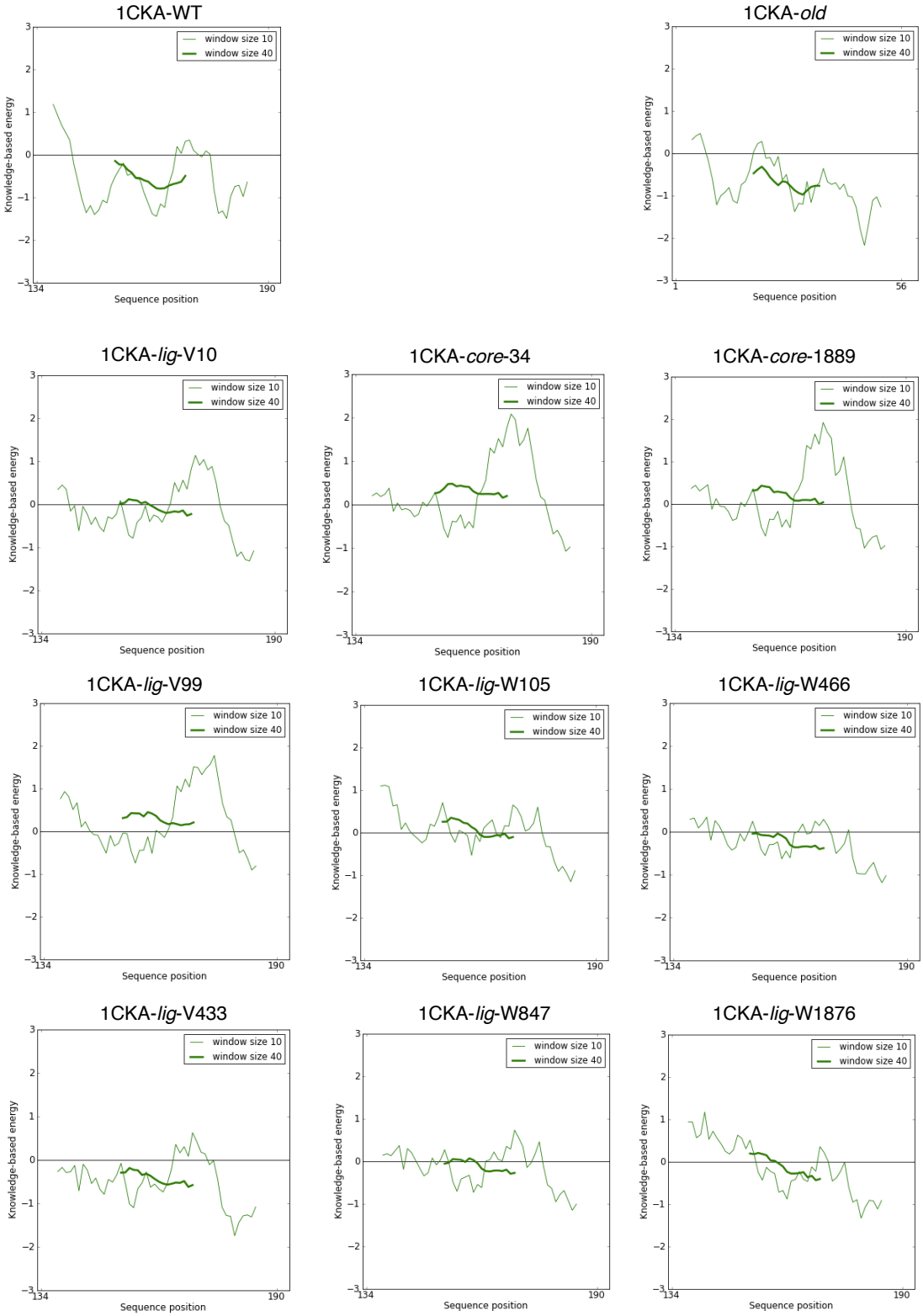


Figure B.15 – Énergies de chaque résidu des séquences des protéines sauvage et mutantes 1CKA par le programme ProsaII.

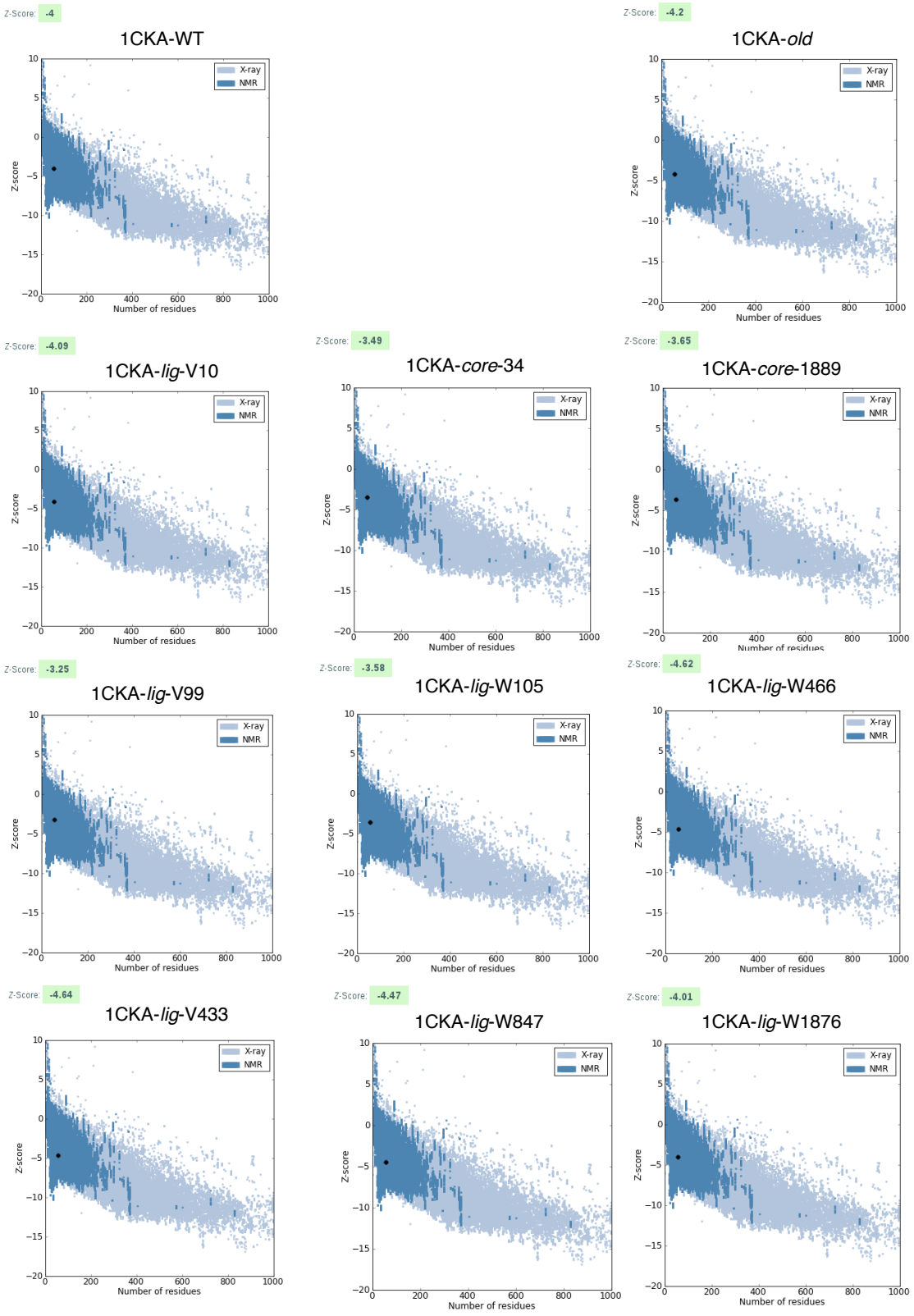


Figure B.16 – Z-scores des protéines sauvage et mutantes 1CKA par le programme ProsaII.

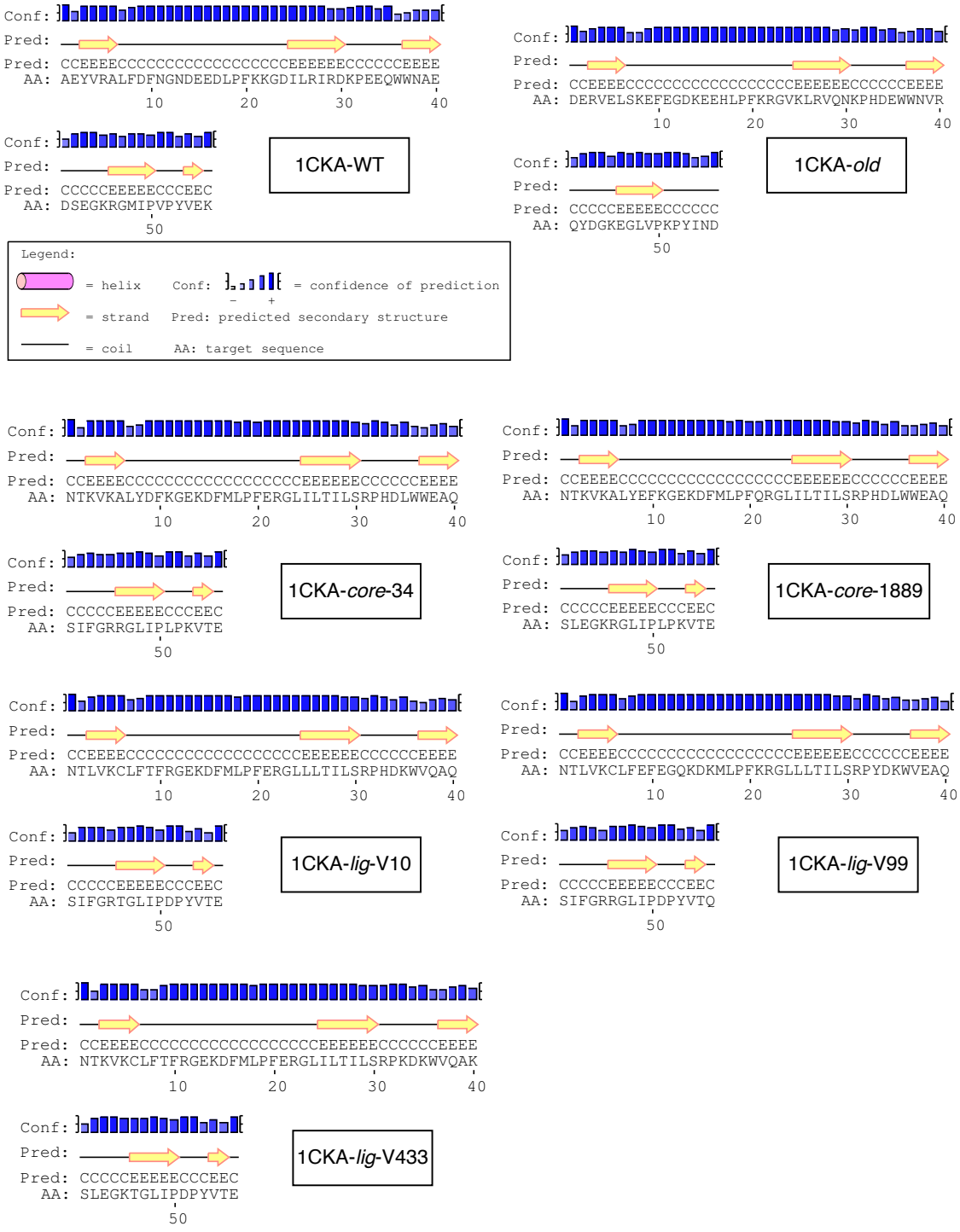


Figure B.17 – Prédiction de structures par le programme Psipred à partir de la séquence native 1CKA-WT, et des séquences mutantes 1CKA-*old* et 1CKA-*LIG* et 1CKA-*CORE* choisies pour l'étude expérimentale.

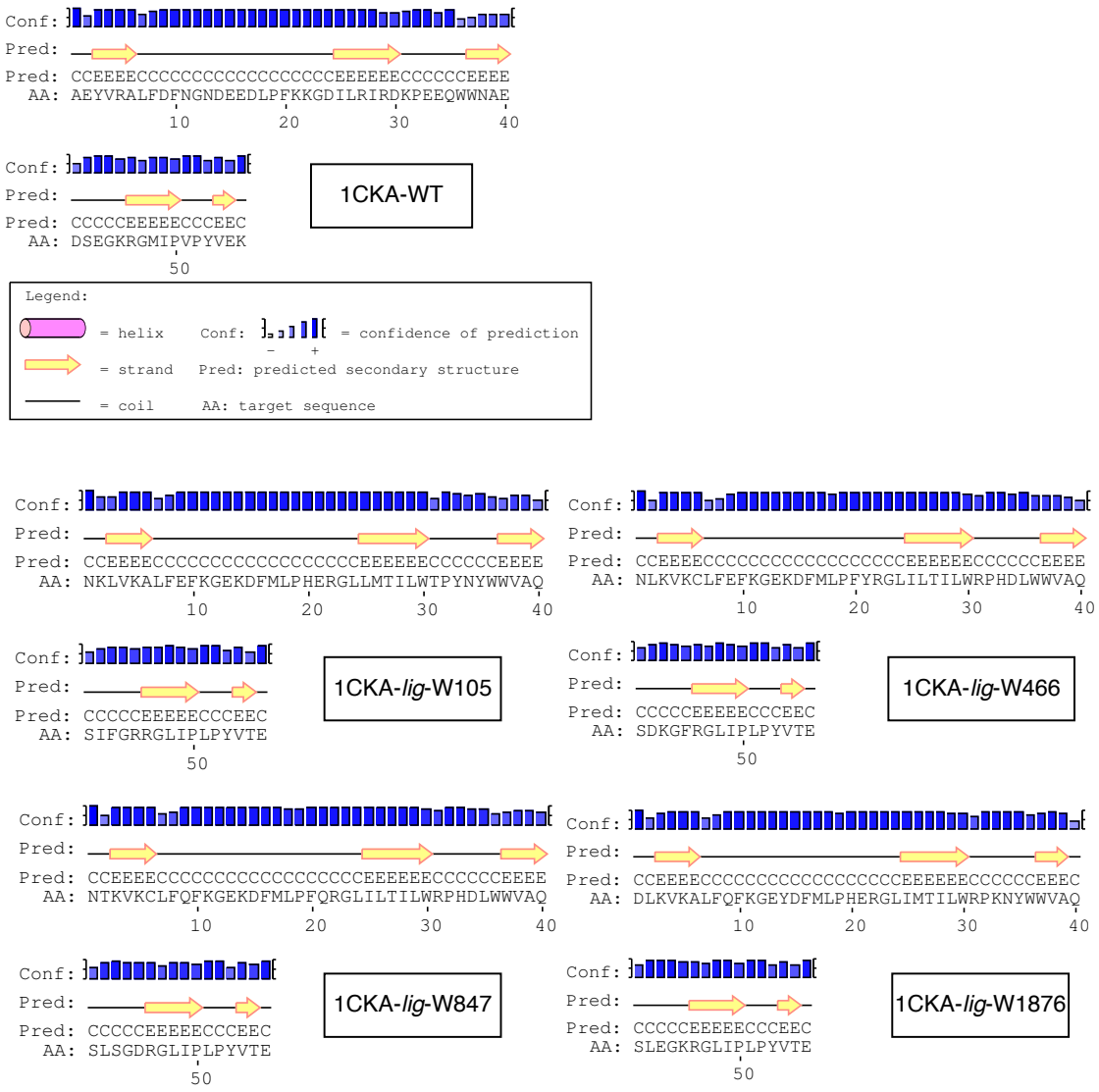


Figure B.18 – Prédiction de structures par le programme Psipred à partir de la séquence native 1CKA-WT, et des séquences mutantes 1CKA-*LIG* choisies pour l'étude expérimentale.

Génération	Aggrescan (TA)	Tango	Aliphatic index	Instability	Gravy
old	-17,79	111,70	59,65	31,60 (s)	-1,398
V10	8,29	154,53	92,28	27,95 (s)	-0,019
V99	5,87	102,15	92,28	46,73 (u)	-0,209
V114	14,61	120,50	111,05	37,89 (s)	0,153
V135	15,05	140,90	97,32	32,27 (s)	0,133
V433	3,76	171,88	85,44	20,64 (s)	-0,274
V1445	9,67	171,62	92,28	34,13 (s)	-0,084
V2857	11,09	122,22	95,61	36,05 (s)	0,130
W105	14,54	113,45	94,04	29,22 (s)	-0,014
W144	14,43	174,03	94,04	35,71 (s)	-0,019
W421	11,17	135,25	99,12	55,82 (u)	-0,091
W466	14,06	132,74	99,12	35,93 (s)	0,047
W589	8,44	117,72	94,04	32,38 (s)	-0,102
W847	11,04	142,98	99,12	33,51 (s)	0,002
W1876	10,06	133,44	94,04	26,60 (s)	-0,181
C34	5,46	118,41	95,79	44,88 (u)	-0,204
C243	7,03	138,09	95,79	34,80 (s)	-0,130
C428	9,41	124,44	95,79	49,98 (u)	-0,019
C1889	2,77	130,84	95,79	34,91 (s)	-0,316
C2469	11,23	122,39	107,72	40,34 (u)	-0,049
L1962	6,26	162,94	75,09	39,32 (s)	-0,425
L2533	-0,24	179,78	71,75	17,44 (s)	-0,533
L4333	0,34	140,70	114,56	42,35 (u)	0,154

Table B.1 – Caractéristiques sur l'aggrégation théorique des protéines mutantes. *Aggrescan* : bioinfo.uab.es/aggrescan/ probabilité d'agrégation. *Tango* : tango.crg.es β-agrégation, calculé à pH 7.0, 298,15K et ionic strength=0,02. Protparam : *aliphatic index*, *Instability* et *Gravy* (grand average of hydropathicity) calculés sur les séquences avec une MET initiatrice.

Protéine	Aggrescan (TA)	Tango	Aliphatic index	Instability	Gravy	pI
1CKA	-14,37	74,08	61,58	57,82 (u)	-1,028	4,54
1CSK	-1,82	137,68	75,26	23,26 (s)	-0,340	7,77
1UTI	-0,673	137,16	82,11	56,69 (u)	-0,244	4,64
1SEM	-4,268	101,56	66,67	39,61 (s)	-0,524	5,04
1ABO	-0,565	125,69	78,60	14,49 (s)	-0,298	5,54

Table B.2 – Caractéristiques sur l'aggrégation théorique des protéines sauvages. *Aggrescan* : bioinfo.uab.es/aggrescan/ probabilité d'agrégation. *Tango* : tango.crg.es β-agrégation, calculé à pH 7.0, 298,15K et ionic strength=0,02. Protparam : *aliphatic index*, *Instability* et *Gravy* (grand average of hydropathicity) calculés sur les séquences avec une MET initiatrice, et point isoélectrique théorique.

Bibliographie

Allen B. & Mayo S. (2012). An efficient algorithm for multistate protein design based on. *J. Comp. Chem.* **31**, 904–916.
cité page 129

Aloy P., Stark A., Hadley C. & Russell R. (2003). Predictions without templates : new folds, secondary structure, and contacts in casp5. *Proteins* **53**, 436–456.
cité page 24

Altschul S., Gish W., Miller W., Myers E. & Lipman D. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–10.
cité page 27

Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W. & Lipman D. (1997). Gapped balst and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
cité page 28

Andersen H.C. (1983). Rattle: a 'velocity' version of the shake algorithm for molecular dynamics calculations. *Journal of Chemical Physics* **52**, 24–34.
cité page 54

Anfinsen C.B. (1973). Structure of a protein is determined solely by the amino acids sequence info. *Science* **181**, 223–230.
cité page 5

Anfinsen C.B., Redfield R.R., Choate W.L., Page J. & Carroll W.R. (1954). Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *The Journal of Biological Chemistry* **207**, 201–210.
cité page 5

Angrand I., Serrano L. & Lacroix E. (2001). Computer-assisted redesign of spectrin sh3 residue clusters. *Biomolecular Engineering* **18(3)**, 125–134.
cité page 232

Archontis G. & Simonson T. (2005). *J. Phys. Chem. B* **109**, 22667–22673.
cité page 48

Baker D. (2006). Prediction and design of macromolecular structures and interactions. *Phil. Trans. R. Soc. Lond.* **361**, 459–463.
cité page 34

Bakowies D. & Gunsteren W. (2002). Water in protein cavities : A procedure to identify internat water and exchange pathways and application to fatty acid-binding protein. *Proteins* **47**, 534–545.
cité page 168

Bar-Sagi D., Rotin D., Betzer A., Mandiyan V. & Schlessinger J. (1993). Sh3 domains direct cellular localisation of signaling molecules. *Cell* **74**, 83–91.
cité page 69

Barlow D. & Thornton J. (1988). Helix geometry in proteins. *J Mol Biol* **201**, 601–19.
cités pages 17 et 18

Bashford D. & Case D. (2000). Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem* **51**, 129–152.
cité page 47

Benson D., Karsch-Mizrachi I., Liman D., Ostell J. & Sayers E. (2009). *Genbank. Nucleic Acids Research* **37**.
cité page 22

Berendsen H., Postma J., van Gunsteren W. & J. H. (1981). Intermolecular forces. *D. Reidel Publishing Company.* •, •.
cité page 45

Berendsen H., Grigera J. & Straatsma T. (1987). The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271.
cité page 45

Berg T. (2003). Modulation of protein-protein interactions with small organic molecules. *Angew Chem Int Ed* **42**, 2462–81.
cité page 76

Berman H., Westbrook J., Feng Z., Gilliland G., Bhat T., Weissig H., Shindyalov I. & Bourne P. (2000). The protein data bank. *Nucleic Acids Res* **28**, 235–242.
cité page 20

Berne B. & Pecora R. (1976). Dynamic light scattering. *Wiley-Interscience, New-York* .
cité page 185

Biou V., Gibrat J., Levin J., Robson B. & Garnier J. (1988). Secondary structure prediction: combination of three different methods. *Protein Eng.* **2**, 185–191.
cité page 9

Bodenhausen G. & Ruben D. (1980). Natural abundance nitrogen 15 nmr by enhanced heteronuclear spectroscopy. *Chem Phys Lett* **69**, 185–9.
cité page 225

Bolon D. & Mayo S. (2001). Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA* **98**, 14274–14279.
cité page 34

Bonneau R., Tsai J., Ruczinski I., Chivian D., Rohl C., Strauss C. & Baker D. (2001). Rosetta in casp4 : progress in ab initio protein structure prediction. *Proteins* **Suppl 5**, 119–26.
cité page 24

Borreguero J., Ding F., Buldyrev S., Stanley H. & Dokholyan N. (2004). Multiple folding pathways of the sh3 domain. *Biophys J* **87**, 521–33.
cité page 73

Bowie J., Luthy R. & Eisenberg D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–70.
cité page 31

Bradley P., Chivian D., Meiler J., Misura K., Rohl C., Schief W., Wedemeyer W., Schueler-Furman O., Murphy P., Schonbrun J., Strauss C. & Baker D. (2003). Rosetta predictions in casp5 : successes, failures, and prospects for complete automation. *Proteins* **53**, 457–68.
cité page 24

Brehelin L., Florent I., Gascuel O. & Marechal E. (2010). Assessing functional annotation transfers with inter-species conserved coexpression : application to plasmodium falciparum. *BMC genomics* **11**, 35.
cité page 22

Brenner S., Chothia C. & Hubbard T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA* **95**, 6073–8.
cités pages 30 et 80

Brooks B., Bruccoleri R., Olafson B., States D., Swaminathan S. & Karplus M. (1983). Charmm : a program for macromolecular energy, minimization, and molecular dynamics calculations. *J. Comp. Chem.* .
cités pages 43 et 184

Broutin I. & Ducruix A. (2000). Domaines structuraux et signalisation. *Medecine Sciences* **16**, 611–616.
cité page 67

Butterfoss G. & Kuhlman B. (2006). Computer-based design of novel protein structures. *Ann. Rev. Biophys. Biomolec. Struct.* **35**, 49–65.
cité page 34

Bystroff C. & Baker D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* **281**, 565–77.
cité page 24

Cartailler J. & Luecke H. (2004). Structural and functional characterization of pi bulges and other short intrahelical deformations. *Structure (Camb)* **12**, 133–144.
cité page 18

Carugo O. & Argos P. (1997). Protein-protein crystal-packing contacts. *Protein Sci* **6**, 2261–2263.
cité page 58

Casares S., Sadqi M., López-Mayorga O., Conejero-Lara F. & van Nuland N. (2004). Detection and characterization of partially unfolded oligomers of the sh3 domain of alpha-spectrin. *Biophysical Journal* **86**, 2403–2413.
cité page 234

Case D., Pearlman D., Caldwell J., Cheatham III T., Ross W., Simmerling C., Darden T., Merz K., Stanton R., Cheng A., Vincent J., Crowley M., Tsui V., Radmer R., Duan Y., Pitera J., Massova I., Seibel G., Singh U., Weiner P. & Kollman P. (1999). Amber 6. *University of California, San Francisco* .
cité page 43

CASP (-). http//predictioncenter.llnl.gov/casp6/casp6.html. - .
cité page 27

Chan A., Hutchinson E., Harris D. & Thornton J. (1993). Identification, classification and anlysis of beta-bulges in proteins. *Protein Sci* **2**, 1574–90.
cité page 18

Chothia C. (1992). Proteins, one thousand families dor the molecular biologist. *Nature* **357**, 543–4.
cité page 82

Cicchetti P., Mayer B., Thiel G. & Baltimore D. (1992). Identification of a protein that binds to the sh3 of abl and is similar to bcr and gap-rho. *Science* **257**, 803–6.
cité page 69

Cohen B., Ren R. & Baltimore D. (1995). Modular binding domains in signal transduction proteins. *Cell* **80**, 237–48.
cité page 69

Conchillo-Sole O., de Groot N., Aviles F., Vendrell J., Daura X. & Ventura S. (2007). Aggrescan: a server efor the prediction and evaluation of hot spots of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65.
cité page 216

Cornell W., Cieplack P., Bayly C., Gould I., Merz K., Ferguson D., Spellmeyer D., Fox T., Caldwell J. & Kollman P. (1995). A second force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* .
cité page 127

Cornette J., Cease K., Margalit H., Spouge J. & Berzofsky JA. ans De-Lisi C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol* **195**, 659–85.
cité page 9

Coulson A. & Moult J. (2002). A unifold, mesofold and superfold model of protein fold use. *Proteins* **46**, 61–71.
cité page 82

Creamer T. & Campbell M. (2002). Determinants of the polyproline ii helix from modeling studies. *Adv Protein Chem* **62**, 263–82.
cité page 74

Crick F. (1953). The packing of alpha-helicies : simple coiled-coils. *Acta. Crystallogr.* **6**, 689–697.
cité page 37

Cusack S., Hartlein M. & Leberman N. (1990). A second class of synthetase structure revealed by xray analysis of escherichia coli seryl-trna synthetase at 2.5 ang. *Nature* **347**, 249–255.
cité page 59

Dahiyat B. & Mayo S. (1996). Protein design automation. *Protein Science* **5**, 895–903.
cités pages 39 et 131

Dahiyat B. & Mayo S. (1997). De novo protein design : fully automated sequence selection. *Science* **278**, 82–87.
cités pages 37 et 59

Dahiyat B., Gordon D. & Mayo S. (1997). Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333–1337.
cité page 43

Dalgarno D., Botfield M. & Rickles R. (1997). Sh3 domains and drug design: ligands, structure, and biological function. *Biopolymers* **43**, 383–400.
cité page 65

Dantas G., Kuhlman B., Callender D., Wong M. & Baker D. (2003a). A large test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460.
cité page 34

Dantas G., Kuhlman B., Callender D., Wong M. & Baker D. (2003b). A large scale test of computational protein design : folding and stability of nine completely redesigned globular proteins. *J Mol Biol* **332**, 449–460.
cité page 24

Dasgupta S., Iyer G., Bryant S., Lawrence C. & Bell J. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins* **28**, 494–514.
cité page 58

Dayhoff M., Eyck R. & Park C. (1972). A model of evolutionary change in proteins. *Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC* **5**, 89–99.
cité page 9

Degterev A., Lugovskoy A., Cardone M., Mulley B., Wagner G., Mitchisonn T. & Yuan J. (2001). Identification of small-molecule inhibitors of interaction betwenn the bh3 domain and bcl-xl. *Nat Cell Biol* **3**, 173–82.
cité page 77

Desjarlais J. & Handel T. (1999). Sidechain and backbone flexibility in protein core design. *J. Mol. Biol.* **289**, 305–318.
cité page 37

Donohue J. (1953). Hydrogen bonded helical configurations of the polypeptide chain. *PNAS* **39**, 470–8.
cité page 17

Dunbrack R. & Karplus M. (1993a). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol* **230**, 543–574.
cité page 24

Dunbrack R. & Karplus M. (1993b). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
cité page 129

Dwyer M., Looger L. & Hellinga H. (2004). Computational design of a biologically active enzyme. *Science* **304**, 1967–71.
cité page 32

Eddy S. (1998). Profil hidden markov models. *Bioinformatics* **14**, 755–763.
cité page 31

Edelsbrunner H., Facello M. & Liang J. (1996). On the definition and the construction of pockets in macromolecules. *Pac. Symp. Biocomput.* 272–287.
cité page 168

Eisenberg D. (1982). A problem for the theory of biological structure. *Nature* **295**, 99–100.
cité page 32

Eswar N., Ramakrishnan C. & Srinivasan N. (2003). Stranded in isolation : structural role of isolated extended strands in proteins. *Protein Eng* **16**, 331–9.
cité page 16

Evans J. (1995). Biomolecular nmr spectroscopy. *Oxford Universeity Press* .
cité page 224

Felsenstein J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.
cité page 9

Feng S., Kapoor T., Shirai F., Combs A. & Schreiber S. (1996). Molecular basis for the binding of sh3 ligands with non peptide elements identified by combinatorial synthesis. *Chem Biol* **3**, 661–70.
cité page 77

Fernandez-Ballester G., Blanco-Mira C. & Serrano L. (2004). The tryptophan switch : changing ligand-binding specificity from type i to type ii in sh3 domains. *J Mol Biol* **335**, 619–29.
cité page 75

Fernandez-Escamilla A., Rousseau F., Schymkowitz J. & Serrano L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol e-pub* .
cité page 217

Filimonov V., Azuagaa A., Viguerac A., Serrano L. & Mateo P. (1999). A thermodynamic analysis of a family of small globular proteins: Sh3 domains. *Biophysical Chemistry* **77**, 195–208.
cité page 234

Fodje M. & Al-Karadaghi S. (2002). Occurence, conformational features and amino acid propensities for the pi-helix. *Protein Eng* **15**, 353–8.
cité page 17

Fortenberry C., Bowman E., Proffitt W., Dorr B., Combs S., Harp J., Mizoue L. & Meiler J. (2011). Exploring symmetry as an avenue to the computational design of large protein domains. *J. Am. Chem. Soc.* **133**, 18026–18029.
cité page 35

Fraternali F. & van Gunsteren W. (1996). An efficient mean solvatation force model for use in molecular dynamics simulation of proteins in aqueous solution. *J Mol Biol* **256**, 939–948.
cité page 133

Fry M., Panayotou G., Booker G. & Waterfield M. (1993). New insights into protein-tyrosine kinase receptor. *Protein Science* **2**, 1785–97.
cité page 69

Gerstein M. & Hegyi H. (2001). Annotation transfer for genomics : measuring functional diver- gence in multi-domain proteins. *Genome Research* **11**, 1632–1640.
cité page 64

Gilson M. & Honig H. (1986). The dielectric constant of a folded protein. *Biopolymers* **25**, 2097–2119.
cité page 47

Gray J. (2006). High-resolution protein-protein docking. *Curr Opin Struct Biol* **16**, 183–193.
cité page 58

Grigoryan G., Kim Y., Acharya R., Axelrod K., Jain R., Willis L., Dmdic M., Kikkawa J. & DeGrado W. (2011). Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* **332**, 1071–1076.
cité page 35

Guérois R. & Lopez de la Paz M. (2007). Protein design: Methods and applications. *Humana Press* .
cité page 34

Hahne F., Mehrle A., Arlt D., Poustka A., Wiemann S. & Beissbarth T. (2008). Ex- tending pathways based on gene lists using interpro domain signatures. *BMC bioinformatics* **9**, 3.
cité page 22

Harbury P., Plecs J., Tidor B., Alber T. & Kim P. (1998). High-resolution protein design with backbone freedom. *Science* 1462–1467.
cité page 37

Hardin C., Pogorelov T. & Luthey-Schulten Z. (2002). Ab initio protein structure prediction. *Curr Opin Struct Biol.* **12**, 176–181.
cité page 23

Havranek J. & Harbury P. (2003). Automated design of specifity in molecular recognition. *Nat. Struct. Mol. Biol.* **10**, 45–52.
cité page 34

Hawkins G., Cramer C.J. & Truhlar D. (1995). Pairwise solute screening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122–129.
cité page 48

Hawkins G., Cramer C.J. & Truhlar D. (1996). Parameterized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **100**, 19824–19839.
cité page 48

Hendlich M., Lackner P., Weitckus S., Floeckner H., Froschauer R., Gottsbacher K., Casari G. & Sippl M. (1990). Identification of native protein folds amongst a large number of incorrect models. the calculation of low energy conformations from potentials of mean force. *J Mol Biol* **216**, 167–70.
cité page 80

Henikoff S. & Henikoff J. (1992). Amino acid substitution matrices from protein blocks. *PNAS* **89**, 10915–10919.
cité page 167

Hill C., Anderson D., Wesson L., DeGrado W. & Eisenberg D. (1990). Crystal structure of alpha1 : Implications for protein design. *Science* **249**, 543–546.
cité page 44

Hirakawa H., Muta S. & Kuhara S. (1999). The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics* **15**, 141–8.
cité page 135

Hockney R. & Eastwood J. (1988). Computer simulation using particles. *Bristol Hilger* .
cité page 56

Hornak V., Abel R., Okur A., Strockbine B., Roitberg A. & Simmerling C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725.
cité page 127

Hutchinson E. & Thornton J. (1994). A revised set of petentials for beta-turn formation in proteins. *Protein Sci.* **3**, 2207–16.
cité page 17

Illergard K., Ardell D. & Elofsson A. (2009). Structure is three to ten times more conserved than sequence a study of structural response in protein cores. *Proteins* **77**, 499–508.
cités pages 31 et 80

Janin J. & Rodier F. (1995). Protein-protein interaction at crystal contacts. *Proteins* **23**, 580–587.
cité page 58

Janin J. & Wodak S. (1978). Conformation of amino acid sidechains in proteins. *J. Mol. Biol.* **125**, 357–386.
cités pages 36 et 129

Jiang L., Althoff E., Clemente F., Doyle L., Rothlisberger D., Zanghellini A., Gallaher J., Betker J., Tanaka F., Barbas C., Hilvert D., Houk K., Stoddard B. & Baker D. (2008). De novo computational design of retro-aldo enzymes. *Science* **319**, 1387–1391.
cité page 34

Jones D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202.
cité page 216

Jones D., Taylor W. & Thornton J. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–282.
cité page 10

Jones S. & Thornton J. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci USA* **93**, 13–20.
cité page 58

Jorgensen W. & Tirado-Rives J. (1988). The opls potential function for proteins, energy minimization for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666.
cité page 43

Jorgensen W., Chandrasekhar J., Madura J., Impey R. & Klein M. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926.
cité page 45

Kabsch W. & SAnder C. (1983). Dictionary or protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637.
cité page 17

Kang H., Freund C., Duke-Cohan J., Musacchio A., Wagner G. & Rudd C. (2000a). Sh3 domain recognition of a proline-independent tyrosine-based rkxxyxxy motif in immune cell adaptor skap55. *The EMBO Journal* **19**, 2889 – 2899.
cité page 235

Kang H., Freund C., Duke-Cohan J., Musacchio A., Wagner G. & Rudd C. (2000b). Sh3 domain recognition of a proline-independent tyrosine-based rkxxyxxy motif in immune cell adaptor skap55. *EMBO J* **19**, 2889–99.
cité page 76

Karplus K., Sjolander K., Barrett C., Cline M., Haussler D., Hughey R., Holm L. & Sander C. (1997). Predicting protein structure using hidden markkov models. *Proteins* **Suppl 1**, 134–139.
cité page 31

Karplus K., Barrett C. & Hughey R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856.
cité page 31

Karplus M. & McCammon J.A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **9**, 646–652.
cités pages 53 et 54

Kawashima S., Ogata H. & Kanehisa M. (1999). Aaindex : Amino acid index database. *Nucleic Acids Res* **27**, 368–9.
cité page 9

Kelly S., Jess T. & Price N. (1751). How to study proteins by circular dichroism. *Biochim Biophys Acta* 119–139.
cité page 221

King N., Scheffer W., Sawaya M., Vollmar B., Sumida J., Andre I., Gonen T., Yeates T. & Baker D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174.
cité page 35

Koga N., Tatsumi-Koga R., Liu G., Xiao R., Acton T., Montelione G. & Baker D. (2012). Principles for designing ideal protein structures. *Nature* **491**, 222–224.
cité page 34

Kono H. & Doi J. (1994). Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins* **19**, 244–255.
cité page 129

Kraemer-Pecore C., Lecomte J. & Desjarlais J. (2003). De novo redesign of the ww domain. *PS* **12**, 2194–2105.
cité page 59

Kuhlman B., Dantas G., Ireton G., Varani G., Stoddard B. & Baker D. (2003a). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.
cité page 34

Kuhlman B., Dantas G., Ireton G., Varani G., Stoddard B. & Baker D. (2003b). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–8.
cités pages 32, 60 et 61

Kumar S. & Bansal M. (1998). Dissecting alpha-helices : position-specific analysis of alpha-helices in globular proteins. *Proteins* **31**, 460–476.
cité page 18

Kuryan J. & Cowburn D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Ann Rev Biomol Struct* **26**, 259–88.
cité page 70

Lanci C., MacDermaid C., Kang S., Acharya R., North B., Yang X., Qiu X., DeGrado W. & Saven J. (2012). Computational design of a protein crystal. *Proc. Natl. Acad. Sci. USA* **109**, 7304–7309.
cité page 35

Leach A. (2001). Molecular modeling : principles and applications. *Pearson Education, Harlow, 2nd edition* .
cité page 53

Lee B. & Richards F. (1971). The interpretation of protein structures : estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
cités pages 50 et 128

Levinthal C. (1969). How to fold graciously? *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois* 22–24.
cité page 24

Levitt M. & Warshel A. (1975). Computer simulation of protein folding. *Nature* **253**, 694–98.
cité page 25

Lewis P., Momany F. & Scheraga H. (1971). Folding of polypeptide chains in proteins ; a proposed mechanism for folding. *PNAS* **68**, 2293–7.
cité page 17

Liang H., Chen H., Fan K., Wei P., Guo X., Jin C., Zeng C., Tang C. & Lai L. (2009). De novo design of a beta-alpha-beta motif. *Ang. Chemie Int. Ed.* **48**, 3301–3303.
cité page 34

Liang J. & Dill K. (2001). Are proteins well-packed ? *Biophys. J.* **81**, 751–766.
cité page 168

Linding R., Diella F., Rousseau F., Schymkowitz J. & Serrano L. (2004). A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* .
cité page 217

Lippow S. & Tidor B. (2007). Progress in computational protein design. *Curr. Opin. Biotech.* **18**, 305–311.
cité page 34

Liu Q., Berry D., Nash P., Pawson T., McGlade C. & Li S. (2003). Structural basis for specific binding of the gads sh3 domain to an rxxk motif-containing slp-76 peptide: a novel mode of peptide recognition. *Mol Cell* **11**, 471–81.
cité page 76

Liu W., Vidal M., Gresh N., Roques B. & Garbay C. (1999). Small peptides containing phosphotyrosine and adjacent alphame-phosphotyrosine or its mimetics as highly potent inhibitors of grb2 sh2 domain. *J Med Chem* **42**, 3737–41.
cité page 78

Liu X., Fan K. & Wang W. (2004). The number of protein folds and their distribution over families in nature. *Proteins* **54**, 491–9.
cité page 82

Looger L., Dwyer M., Smith J. & Hellinga H. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185–190.
cité page 34

Lopes A., A. A., Bathelt C., Archontis G. & Simonson T. (2007). Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins* **67**, 853–67.
cité page 156

Lopes A., Schmidt am Busch M. & Simonson T. (2009). Computational design of protein :ligand binding : modifying the specificity of asparaginyl-trna synthetase. *journal JCC* .
cité page 59

Lovell S., Word J., Ridchardson J. & Ridchardson D. (1999). Asparagine and glutamine rotamers : B-factor cutoff and correction of amide flips yield distinct clustering. *Proc. Natl. Acad. Sci. USA* **96**, 400–405.
cité page 129

Low B. & Baybutt R. (1952). The pi-helix : a hydrogen bonded configuration of the polypeptide chain. *J Am Chem Soc* **74**, 5806.
cité page 17

Lowenstein E., Daly R., Batzer A., Li W., Margolis B., Lammers R., Ullrich A., Skolnik E., Bar-Sagi D. & Schlessinger J. (1992). The sh2 and sh3 domains-containing protein grb2 links receptor tyrosine kinase to ras signaling. *Cell* **70**, 431–2.
cité page 69

Lumb K. & Kim P. (1995). A buried interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**, 8642–8648.
cité page 44

Maiorov V. & Crippen G. (1994). Significant of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* **235**, 625–634.
cité page 186

Malakauskas S. & Mayo S. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475.
cité page 36

Marin A., Pothier J., Zimmermann K. & Gibrat J. (2002). Frost : a filter-based fold recognition method. *Proteins* **49**, 493–509.
cité page 31

Marshall S. & Mayo S. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**, 619–631.
cité page 59

McCammon J.A., Gelin B.R. & Karplus M. (1977). Dynamics of folded proteins. *Nature* **267**, 585–590.
cité page 53

McGregor M., Islam S. & Sternberg M. (1987). Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295–310.
cité page 129

Mendez R., Leplae R., Lensink M. & Wodak S. (2005). Assessment of capri predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **60**, 150–169.
cité page 58

Morel B., Casares S. & Conejero-Lara F. (2006). A single mutation induces amyloid aggregation in the alpha-spectrin sh3 domain: Analysis of the early stages of fibril formation. *Journal of Molecular Biology* **356**, 453–468.
cité page 234

Murzin A., Lesk A. & Chothia C. (1994). Principles determining the structure of beta-sheet barrels in proteins. i. a theoretical analysis. *J. Mol. Biol.* **236**, 1369–1381.
cité page 37

Needlman S. & Wunsch C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–53.
cité page 26

Nguyen J., Turck C., Cohen F., Zuckermann R. & Lim W. (1998). Exploiting the basis of proline recognition by sh3 and ww domains: design of n-substituted inhibitors. *Science* **282**, 2088–92.
cité page 77

Nguyen J., Porter M., Amoui M., Miller W., Zuckermann R. & Lim W. (2000). Improving sh3 domain ligand selectivity using a non-natural scaffold. *Chem Biol* **7**, 463–73.
cité page 77

Nicholls P. (2000). Introduction : the biology of the water molecule. *Cell. Mol. Life Sci.* **57**, 987–992.
cité page 44

Ockey D. & Gadek T. (2002). Inhibitors of protein-protein interactions. *Expert Opin Ther Patents* **12**, 393–400.
cité page 76

Ohnishi S., Lee A., Edgell M. & Shortle M. (2004). Direct demonstration of structural similarity between native and denaturated eglin c. *Biochemistry* **43**, 4064–4070.
cité page 39

Okabe A., Boots B., Sugihara K. & Chiu S. (2000). Spatial tessellations : concepts and applications of voronoi diagrams. *Wiley series in probability and statistics. John WWiley and Sons* .
cité page 168

Oldfield E. (1995). Chemical shifts et three-dimensional protein structures. *J Biomol NMR* **5**, 217–25.
cité page 225

Oneyama C., Nakano H. & Sharma S. (2002). Ucs15a, a novel small molecule, sh3 domain-mediated protein-protein interaction blocking drug. *Oncogene* **21**, 2037–50.
cité page 77

Oneyama C., Agatsuma T., Kanda Y., Nakano H., Sharma S., Nakano S., Narazaki F. & Tatsuta K. (2003). Synthetic inhibitors of proline-rich ligand- mediated protein-protein interaction: potent analogs of ucs15a. *Chem Biol* **10**, 443–51.
cité page 77

Opatowsky Y., Chomsky-Hecht O., Kang M.G., Campbell K. & Hirsch J. (2003). The voltage-dependent calcium channel beta subunit contains two stable interacting domains. *The Journal of Biological Chemistry* **278**, 52323–52332.
cité page 232

Pagliaro L., Felding J., Audouze K., Nielsen S., Terry R., Krog-Jensen C. & Butcher S. (2004). Emerging classes of protein-protein interaction inhibitors and new tools for their development. *Curr Opin Chem Biol* **8**, 442–9.
cité page 76

Pande V. & Rokhsar D. (1998). Is the molten globule a third phase of proteins? *Proc Natl Acad Sci USA* **95**, 1490–4.
cité page 239

Pantazes R., Greenwood M. & Maranas C. (2011). Recent advances in computational protein design. *Curr. Opin. Struct. Biol.* **21**, 467–472.
cité page 34

Pauling L. & Corey R. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci USA* **37**, 251–6.
cité page 13

Pauling L., Corey R. & Branson H. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* **37**, 205–11.
cité page 13

Pawson T. (1995). Protein modules and signalling networks. *Nature* **373**, 573–80.
cité page 69

Peters K., Fauck J. & Frommel C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.
cité page 168

Peterson R., Dutton P. & Wand A. (2004). Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Prot. Sci.* **13**, 735–751.
cité page 129

Pleiss J. (2011). Protein design in synthetic biology. *Curr. Opin. Biotech.* **22**, 611–617.
cité page 34

Pokala N. & Handel T. (2005). Energy functions for protein design : Adjustement with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **347**, 203–227.
cité page 39

Ponder J. & Richards F. (1987). Tertiary templates for proteins : use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–792.
cités pages 36 et 129

Punta M., Coggill P., Eberhardt R., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E., Eddy S., Bateman A. & Finn R. (2012). The pfam protein families database. *Nucleic Acids Research* **40**, 290–301.
cité page 156

Qi Y., Sadreyev R., Wang Y., Kim B.H. & Grishin N. (2007). A comprehensive system for evaluation of remote sequence similarity detection. *BMC Bioinformatics* **8**, 314.
cité page 30

Ramachandran G. & Sasisekharan V. (1968). Conformation of polypeptides and proteins. *Adv Protein Chem* **23**, 283–438.
cité page 12

Richards F. (1974). The interpretation of protein structures : total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1–14.
cité page 168

Richardson J. & Richardson D. (2002). Natural beta-sheet proteins use negative desiegn to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* **99**, 2754–9.
cité page 18

Richardson J., Getzoff E. & Richardson D. (1978). The beta bulge : a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci USA* **75**, 2574–8.
cité page 18

Richter F., Leaver-Kay A., Khare S., Bjelic S. & Baker D. (2011). De novo enzyme design using rosetta3. *PLoS One* **6**, e19230.
cités pages 35 et 61

Romero F., Dargemont C., Pozo F., Reeves W., Camonis J., Gisselbrecht S. & Fischer S. (1996a). p95vav associates with the nuclear protein ku-70. *Mol Cell Biol* **16**, 37–44.
cité page 76

Romero F., Dargemont C., Pozo F. Reeves W., Camonis J., Gisselbrecht S. & Fischer S. (1996b). P95 vav associates with the nuclear ku-70. *Mol Cell Biol* **16**, 37–44.
cité page 70

Rost B. (1999). Twilight zone ofprotein sequence alignments. *Protein Eng* **12**, 85–94.
cités pages 30 et 80

Rothlisberger D., Khersonsky O., Wollacott A., Jiang L., DeChancie J., Betker J., Gallaher J., Althoff E., Zanghellini A., Dym O., Albeck S., Houk K., Tawfik D. & Baker D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195.
cité page 34

Rousseau F., Schymkowitz J. & Serrano L. (2006). Protein aggregation and amyloidosis: confusion of the kinds? *Current Opinion in structural biology* **16**, 1–9.
cité page 216

Ryckaert J.P., Ciccotti G. & Berendsen H.J.C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Chemical Physics* **23**, 327–341.
cité page 54

Saksela K. & Permi P. (2012). Sh3 domain ligand binding : What's the consensus and where's the specificity? *FEBS Lett* **586**, 2609–14.
cité page 63

Sali A. & Blundell T. (1993). Comparative protein modelling by satifaction of spatial restraints. *J. Mol Biol* **234**, 779–815.
cité page 28

Samish I., MacDermaid C., Perez-Aguilar J. & Saven J. (2011). Theoretical and computational protein design. *Ann. Rev. Phys. Chem.* **62**, 129–149.
cité page 34

Sander C. & Schneider R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68.
cités pages 25, 30 et 80

Saunders C. & Baker D. (2005). Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol.* **346**, 631–644.
cité page 60

Saven J. (2010). Computational protein design: Advances in the design and redesign of biomolecular nanostructures. *Curr. Opin. Colloid Interf. Sci.* **15**, 13–17.
cité page 35

Saven J. (2011). Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. *Curr. Opin. Chem. Biol.* **15**, 452–457.
cité page 34

Scaife R. & Margolis R. (1997). The role of the ph domain and sh3 binding domains in dynamin function. *Cell Signal* **9**, 395–401.
cité page 69

Schlich T. (2002). Molecular modeling and simulations : an interdisciplinary guide. *Springer-Verlag, New-York, USA* .
cité page 53

Schmidt am Busch M., Lopes A., Amara N., Bathelt C. & Simonson T. (2008a). Testing the coulomb/accessible surface area solvent model for protein stability, ligand binding, and protein design. *BMC Bioinformatics* **9**, 148.
cité page 133

Schmidt am Busch M., Lopes A., Mignon D. & Simonson T. (2008b). Computational protein design : software implementation, parameter optimization, and performance of a simple model. *J. Comput. Chem.* **29**, 1092–1102.
cité page 131

Schmidt am Busch M., Sedano A. & Simonson T. (2010). Computational protein design : Validation and possible relevance as a tool for homology searching and fold recognition. *PLoS One* **5**, e10410.
cité page 94

Sean R.E. (2004). Where did the blosum62 alignment score matrix come from ? *Nature Biotechnology* **22**, 1035.
cité page 167

Shangary S. & Johnson D. (2002). Peptides derived from bh3 domains of bcl-2 family members: a comparative analysis of inhibition of bcl-2, bcl- x(l) and bax oligomerization, induction of cytochrome c release, and activation of cell death. *Biochemistry* **41**, 9485–95.
cité page 77

Sigrist C., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A. & Bucher P. (2002). Prosite: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**, 265–74.
cité page 72

Simmerling C., Strockbine B. & Roitberg A. (2002). All-atom structure prediction and folding simulations of a stable protein. *J Am Chem Soc* **124**, 11258–9.
cité page 25

Simons K., Kooperberg C., Huang E. & Baker D. (1997). Assembly of protein tertiaru structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol* **268**, 209–225.
cités pages 24 et 61

Simons K., Bonneau R., Ruczinski I. & Baker D. (1999a). Ab initio protein structure prediction of casp iii tagets using rosetta. *Proteins* 171–176.
cité page 24

Simons K., Ruczinski I., Kooperberg C., Fox B., Bystroff C. & Baker D. (1999b). Improved recognition of native-like protein structures using a combination of sequence dependent and sequence independent features of proteins. *Proteins* **34**, 82–95.
cité page 24

Sippl M. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859–83.
cité page 80

Smith T. & Waterman M. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195–7.
cité page 26

Smithgall T. (1995). Sh2 and sh3 domains : potential targets for anti-cancer drug design. *J Pharmacol Toxicol Methods* **34**, 125–32.
cité page 70

Spiegel K., Degrado W. & Klein M. (2006). Structural and dynamique properties of manganese catalese and the synthetic protein df1 and their implication for reactivity from classical molecular dynamics calculations. *Proteins* **65**, 317–330.
cité page 185

Still W., Tempczyk A., Hawley R. & Hendrickson T. (1990). Semianalytical tratment of solvataion for molecular mechanics and dynamics. *J Am Chem Soc* **112**, 6127–29.
cité page 47

Streett W.B., Tildesley D.J. & Saville G. (1978). Multiple time step methods in molecular dynamics. *Molecular Physics* **35**, 639–648.
cité page 54

Strynadka N., Eisenstein M., Katchalski-Katzir E., Shoichet B., Kunt I., Abagyan R., Totrov M., Janin J., Cherfils J., Zimmerman F., Olson A., Duncan B., Rao M., Jackson R., Sternberg M. & James M. (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to tem-1 beta-lactamase. *Nat Struct Biol* **3**, 233–239.
cité page 58

Su A. & Mayo S. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Prot. Sci.* **9**, 1701–1707.
cités pages 37 et 129

Susva M., Missbach M. & Green J. (2000). Src inhibitors: drugs for the treatment of osteoporosis, cancer or both? *Trends Pharmacol Sci* **21**, 489–95.
cité page 77

Swindells M. (1995). A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.* **4**, 93–102.
cité page 135

Swope W., Anderson H., Berens P. & Wilson K. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules : Application to small water clusters. *J. Chem. Phys.* **76**, 637–49.
cité page 56

Tanaka J., Yangawa H. & Doi N. (2011). Comparison of the frequency of functional sh3 domains with different limited sets of amino acids using mrna display. *PloS ONE* **6**, e18034.
cité page 232

Taylor W. (1986). The classification of amino acid conservation. *J Theor Biol* **119**, 205–218.
cités pages 8 et 9

Tozzini V. (2005). Coarse-grained models for proteins. *Current opinion in structural biology* **15**, 144–150.
cité page 185

Tsui V. & Case D. (2000). Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* **56**, 275–291.
cité page 48

Tuffery P., Etchebest C., Hazout S. & Lavery R. (1991). A new approach to the rapid determination of protein side chain conformations. *Jour. Bio. Str. Dyn.* **8**, 1267–1289.
cité page 127

Tuffery P., Etchebest C. & Hazout S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Engineering* **10**, 361–372.
cité page 127

Vajda S., Vakser I., Sternberg M. & Janin J. (2002). Modeling of protein interactions in genomes. *Proteins* **47**, 444–446.
cité page 58

Vaynberg J. & Qin J. (2006). Weak protein–protein interactions as probed by nmr spectroscopy. *Trends in biotechnology* **24**, 22–27.
cité page 235

Verlet L. (1967). Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev* **159**, 98.
cité page 56

Vidal M., Montiel J., Cussac D., Cornille F., Duchesne M., Parker F., Tocque B., Roques B. & Garbay C. (1998). Differential interactions of the growth factor receptor-bound protein 2 n-sh3 domain with son of sevenless and dynamin. potential role in the ras-dependent signaling pathway. *J Biol Chem* **273**, 5343–8.
cité page 76

Vidal M., Gigoux V. & Garbay C. (2001). Sh2 and sh3 domains as targets for anti-proliferative agents. *Crit Rev Oncol Hematol* **40**, 175–86.
cité page 65

Vidal M., Liu W., Lenoir C., Salzmann J., Gresh N. & Garbay C. (2004). Design of peptoid analogue dimers and measure of their affinity for grb2 sh3 domains. *Biochemistry* **43**, 7336–44.
cité page 78

Voet D. & Voet G. (1998). Biochemistry, second ed. *John Wiley and Sons Inc.* .
cité page 224

Walensky L., Kung A., Escher I., Malia T., Barbuto S., Wright R., Wagner G., Verdine G. & Korsmeyer S. (2004). Activation of apoptosis in vivo by a hydrocarbon-stapled bh3 helix. *Science* **305**, 1466–70.
cité page 77

Wang J., Cieplak P. & Kollman P. (2000). How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J Comput. Chem.* **21**, 1049–1074.
cité page 127

Wang Z. (1996). How many fold types of protein are there in nature ? *Proteins* **26**, 186–91.
cité page 82

Weiner S., Kollman P., Case D., Singh U., Ghio C., Alagona G., Profeta S. & Weiner P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal American Chemical Society* **106**, 765–784.
cité page 127

Wernisch L., Hery S. & Wodak S. (2000). Automatic protein design with all atom force fields by exact and heuristic optimization. *J. Mol. Biol.* **301**, 713–736.
cités pages 39 et 128

Wiederstein & Sippl M. (2007). Prosa-web : interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **35**, W407–W410.
cité page 216

Wolf H. & Haken H. (2004). Molecular physics and elements of quantum chemistry : introduction to experiments and theory. *Springer* .
cité page 224

Wolf Y., Grishin N. & Koonin E. (2000). Estimating the number of protein folds and families from complete genome data. *J Mol Biol* **299**, 897–905.
cité page 82

Wu X., Knudsen B., Feller S., Zheng J., Sali A., Cowburn D., Hanafusa H. & Kuriyan J. (1995). Structural basis for the specific interaction of lysine-containing proline-rich peptides with the n-terminal sh3 domain of c-crk. *Structure* **3**, 215–226.
cité page 228

X-plor version 3.1 A.s.f.X.r.c. & NMR. (1992). Brünger, at. *Yale University Press, New Haven* .
cité page 127

Xia B., Tsui V., Case D., Dyson H. & Wright P. (2002). Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized born model, and with explicit water. *J Biomol NMR* **22**, 317–31.
cité page 48

Xiang Z. & Honig B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311**, 421–430.
cité page 129

Yang Y., Garbay C., Duchesne M., Cornille F., Jullian N., Fromage N., Tocque B. & Roques B. (1994). Solution structure of gap sh3 domain by 1h nmr. *EMBO J* **13**, 1270–9.
cité page 71

Ye Y. & Godzik A. (2004). Comparative analysis of protein domain organization. *Genome Research* **14**, 343–353.
cité page 65

Zarrinpar A., Bhattacharyya R. & Lim W. (2003). The structure and function of proline recognition domains. *Sci STKE* .
cité page 74

Zhang Y., Hubner I., Arakaki A., Shakhnovich E. & Skolnick J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proc Natl Acad Sci USA* **103**, 2605–10.
cité page 82

Résumé

Grâce aux récents progrès technologiques et à l'arrivée des séquenceurs de nouvelle génération, la quantité de données génomiques croît exponentiellement, alors que l'écart avec le nombre de structures résolues se creuse. Dans l'idéal, on aimerait pouvoir prédire par informatique la structure 3D de n'importe quelle protéine à partir de l'information de séquence seule, même en l'absence d'homologie. En effet, en dessous de 30% d'identité de séquence, les mesures de similarité de séquences ne sont plus suffisantes pour détecter l'homologie. Il faut donc mettre en place d'autres méthodes afin de venir à bout de cette zone d'ombre.

Pour une structure donnée (et donc une fonction biologique), on ne dispose souvent que d'une petite quantité de séquences natives y correspondant, et parfois assez peu identiques. Il est alors difficile de construire un profil de recherche d'homologues pour retrouver ces séquences dont on ne connaîtrait pas la structure. Alors comment disposer de bases de données de séquences plus conséquentes pour chaque structure ? Ainsi, le design computationnel de protéine (CPD) tente de répondre à cette problématique : si l'on connaît un repliement, est-il possible de retrouver l'ensemble des séquences qui lui correspondent ?

Le principe du CPD consiste à identifier parmi toutes les séquences compatibles avec le repliement d'intérêt, celles qui vont conférer à la protéine, la fonction désirée. La procédure générale est réalisée en deux étapes. La première consiste à calculer une matrice d'énergie contenant les énergies d'interactions entre toutes les paires de résidus de la protéine en autorisant successivement tous les types d'acides aminés dans toutes leurs conformations possibles. La seconde étape, ou "phase d'optimisation", consiste à explorer simultanément l'espace des séquences et des conformations afin de déterminer la combinaison optimale d'acides aminés étant donné le repliement de départ.

Une première phase d'analyse de covariances de positions d'alignements de séquences théoriques a été menée. Nous avons ainsi pu mettre au point une méthode statistique pour repérer des ensembles de positions qui muteraient ensemble pour une structure donnée. La construction d'un profil avec toutes ces séquences théoriques moyennant trop l'information en acides aminés, nous avons pu améliorer la recherche d'homologues en construisant plusieurs profils à partir de groupes de séquences classées grâce à des motifs sur ces positions considérées comme covariantes.

Pour mieux appréhender la qualité de ces prédictions de séquences théoriques, il fallait mettre en place un protocole de sélection des meilleurs protéines mutantes afin de les tester in vivo. Mais comment déterminer qu'une séquence théorique est meilleure qu'une autre ? Sur quels critères se baser pour les caractériser ? Aussi, un ensemble de descripteurs a été choisi, permettant de trier sur plusieurs critères les séquences théoriques pour n'en choisir qu'une vingtaine. Ensuite, ces protéines mutantes ont été soumises à des simulations de dynamique moléculaire afin d'évaluer leur stabilité théorique. Pour quelques protéines mutantes plus prometteuses, nous avons réalisé des expériences de sur-expression, de purification et de détermination structurale, tentant d'obtenir une validation biologique du modèle de CPD.

Ces protocoles d'analyse et de validation semblent être de bons moyens permettront à notre équipe de tester d'autres protéines mutantes dans l'avenir. Ils pourront ainsi modifier des paramètres lors de la génération par CPD et s'appuyer sur des résultats expérimentaux pour les ajuster.

Mots clés : *design computationnel de protéine, prédiction de structure, recherche d'homologues, dynamique moléculaire, domaines SH3*

Abstract

Computational protein design for structure prediction. Thanks to recent technological breakthroughs and the arrival of new generation sequencers, the amount of genomic data raises exponentially while the gap with the number of solved structures is widening. Ideally, computational 3D structure prediction should be possible with the only sequence information, even without any homology. Indeed, below 30% of sequence identity, similarity measurements are not efficient enough to detect homology. Therefore, it is necessary to implement new methods to take apart the twilight zone.

Usually, for a given structure (and so a biological function), only a few existing sequences is known, and barely similar. Thus it is difficult to build a profile in order to find homologues without knowledge of the structure. How can we have databases of sequences for each structure ? The Computational Protein Design (CPD) try to answer this issue : if a fold is known, it is possible to predict every matching sequence ?

The CPD consists of recognizing, among all compatible sequences with the wanted fold, those whom will confer to the protein the wanted function. Two steps are needed. The first one consists of calculating some energy matrix holding interaction energies between every pair of residues of the protein by allowing successively all types of amino acids in every possible conformation. The second one, or "optimization step", consists of exploring simultaneously spaces of sequences and conformations in order to determine the best combination of amino acids with the fold given at the beginning.

First, the analysis of covariances of alignment positions of theoretical sequences has been managed. We succeeded in the implementation of a statistical method to locate positions that mutate together for a given structure. The profile built with all these theoretical sequences averages too strongly the amino acids data. That is why we improve the homologues searching using groups of sequences classified with the help of patterns located on these positions of covariance.

To appreciate the quality of these predictions of theoretical sequences, we had to implement a selection protocol of the best mutated proteins in order to test them *in vivo*. Nonetheless how can we determine that a sequence is better that another ? What are the relevant criteria ? Thus, a set of descriptors have been chosen to sort the theoretical sequences on the basis of various criteria. Eventually, we got a dozen of sequences. Then, theses mutated proteins have been submitted to molecular dynamics simulations to assess their theoretical stability. For the most encouraging mutated proteins, experimentations took place to get a biological validation of the CPD model : over-expression, purification, structural determination...

These protocols of analysis and validation seem to be good means will allow our team to test other mutant proteins in the future. So they can modify parameters during the generation by CPD and lean on experimental results to adjust them.

Key-words : *computational protein design, structure prediction, homology searching, molecular dynamics, SH3 domain*